Treball de Fi de Grau

Grau en Enginyeria en Tecnologies Industrials

Aplicació de tècniques de mostreig per millorar el rendiment de models de predicció de resultats acadèmics

MEMÒRIA

Autor: Clàudia Figueras Vall
Director: Luis José Talavera Mendez

Convocatòria: Gener 2020

ETSEIB

Escola Tècnica Superior
d'Enginyeria Industrial de Barcelona

UPC

RESUM

El següent document tracta sobre l'aplicació de tècniques de mostreig per tant de millorar el rendiment dels models de predicció de l'aprovat o el suspès dels i les alumnes de l'Escola Industrial de Barcelona de les assignatures del tercer quadrimestre del Grau en Enginyeria en Tecnologies Industrials.

Al llarg del treball s'empren eines que són configurades mitjançant el llenguatge de programació Python, en l'entorn d'Anaconda con a IDE. En concret s'ha utilitzat la llibreria de Pandas i la llibreria de Scikit-learn.

Tot el procediment del treball gira entorn a la metodologia CRISP-DM, des de l'inici amb la preparació de les dades fins al final a la validació dels models. A partir dels resultats obtinguts, s'ha realitzat un contrast entre els resultats d'ambdós mètodes de predicció emprats, amb i sense *oversampling*.

Finalment, la conclusió principal extreta del projecte és que les precisions de predicció són relativament baixes mitjançant l'ús de l'arbres de decisió, tant amb *oversampling* com sense. Això s'atribueix a que les dades que es disposen són poc representatives per tal de poder-ne obtenir una predicció.

SUMARI

1. GLOSSARI

***Anaconda**:* Distribució lliure i oberta de Python i R. Permet el processament de grans volums de informació, anàlisi predictiu i computacions científiques.

DataFrame (DF): Element amb què treballa la llibreria Pandas, representa una taula.

Dataset: Conjunt de dades.

Float: Nombre decimal en llenguatge de Python.

Imbalanced Dataset: Conjunt de dades que, en la variable a predir, presenta més d'un valor que de l'altre.

Int (integer) : Nombre enter en llenguatge de Python

Missing Value: Valors que prenen les cel·les d'un DataFrame quan no tenen cap valor assignat.

NaN (Not a number): Format que prenen els missing vàlues, valors buits.

Pandas: Llibreria de Python emprada durant el treball per al tractament i l'anàlisi de dades.

Python: Llenguatge de programació emprat en el treball.

Random: Aleatori.

Resampling: Tornar a mostrejar.

String: En codi Python, és un element que permet incorporar caràcters tan alfabètics com numèrics. S'inclou entre dues cometes.

Testing: Fase de predicció que inclou la validació dels models estudiats. El conjunt de dades anomenat testing és aquell emprat per dur a terme aquesta acció.

Training: Fase de predicció que inclou la construcció dels models de predicció. El conjunt de dades anomenat training és aquell emprat per dur a terme aquesta acció.

2. INTRODUCCIÓ

Al llarg dels anys, la digitalització de molts processos ha permès facilitar i agilitzar tot tipus de serveis. Això és gràcies a que la tecnologia s'ha anat consolidant com una part essencial de la vida quotidiana, que ha suposat grans canvis en procediments tant senzills com complexos.

Avui en dia, darrera de qualsevol procés s'hi troba una recollida massiva de informació, cosa que representa una generació diària massiva de dades, les quals poden aportar nous coneixements si són gestionades i analitzades.

En aquest tractament és on neix la mineria de dades, que consisteix en l'estudi o anàlisi de dades mitjançant tècniques automàtiques com serien els arbres de decisió, o mètodes que involucren l'acció humana, com per exemple en la identificació de patrons mitjançant computadors. La mineria de dades, és per tant una poderosa eina en l'anàlisi de dades que aplica la tecnologia en la creació de nova informació en forma de identificació de patrons o tendències, així com la detecció d'anomalies i el reconeixement de relacions entre factors.

El projecte es centra en la metodologia CRISP (*Cross-Industry Standard Process*) ja que aquesta és una de les més emprades en la mineria de dades.

2.1. Objectius del projecte

El present treball consisteix en l'anàlisi del rendiment d'una de les tècniques de mineria, els arbres de decisió, en la predicció de l'aprovat o suspès de l'alumnat del grau en Enginyeria en Tecnologies Industrials de l'Escola Tècnica Superior de Barcelona en les assignatures corresponents al tercer quadrimestre, posterior a la fase inicial. Els objectius són els següents:

- **Correcta aplicació d'una metodologia**. Tot el procés d'anàlisi es durà a terme mitjançant una metodologia rigorosa, la qual ha de ser documentada per tal de poder ser replicada en un futur.

- **Estudi i comparació dels resultats obtinguts** amb l'aplicació de tècniques de mostreig o sense, mitjançant l'ús d'arbres de decisió com a mètode de classificació.

- **Validació dels resultats de forma sistemàtica**. En validar el model de predicció obtingut, s'emprarà un mode per tal de realitzar-ne la validació. S'aplicarà el mateix mètode de validació tant en els models predictius en que s'hagi aplicat tècniques de

mostreig com els que no se'ls hagi aplicat.

- **Coneixement de la llibreria Pandas**, amb què es treballarà durant tot el projecte, i **familiarització amb l'entorn**. La llibreria Pandas requereix un cert grau de coneixement per tal de poder implementar les funcions necessàries per al codi informàtic, el qual es durà a terme amb l'IDE Anaconda, ja que aquest presenta totes les eines necessàries per tal de poder realitzar amb èxit el projecte.

2.2. Abast del projecte

Al tractar-se d'un projecte de mineria de dades amb la posterior aplicació de tècniques de mostreig per tal d'incrementar-ne el rendiment, es seguirà la metodologia CRISP, la qual s'exposa amb més detall en el projecte.

L'abast del projecte comprèn totes les fases d'un projecte convencional de mineria de dades, excepte la fase d'implementació, ja que aquesta requeriria imposar el funcionament d'un mètode predictiu a l'escola.

2.3. Eines utilitzades

Totes les eines emprades en el treball són informàtiques. A continuació s'exposen més detalladament.

2.3.1. Python

Python és un llenguatge de programació concebut i implementat a finals dels anys 80, tot i que el seu ús ha anat incrementant de forma exponencial fins el dia d'avui. És un llenguatge característic per la seva simplicitat, alhora que ofereix una gran potència en quant a varietat de possibilitats d'implementació.

El fet que les estructures que permet construït presentin un aspecte molt visual, fa que aquest s'imposi sobre altres codis. A més a més, inclou també una gran varietat de llibreries útils per a tot tipus d'àrees d'estudi.

Segons el portal online KDNuggets, Python és un dels llenguatges més utilitzats, sobretot en àmbits d'intel·ligència artificial i anàlisis predictius. Les enquestes realitzades pel portal a

desenvolupadors d'anàlisi de dades revelen que gran part d'ells opta per l'ús de Python i els elements relacionats amb aquest. Entre les eines més utilitzades també es troben Anaconda, un entorn de desenvolupament per a Python i *scikit-learn*, una llibreria de Python.

S'ha escollit aquest tipus de llenguatge principalment per la gran comunitat d'usuaris que el coneix, la claredat que presenta la lectura del codi i la disposició de llibreries d'anàlisi de dades que ofereix. Cal afegir també, que Python és el llenguatge impartit per la docència realitzada a l'Escola en les assignatures de Fonaments d'Informàtica i Informàtica.

2.3.2. Pandas

Pandas és una llibreria de Python de codi obert, la qual conté tot tipus de funcions que faciliten la manipulació de grans volums de dades. Aquesta llibreria proporciona en especial funcions necessàries per a la preparació de dades. D'aquesta llibreria se'n destaca el següent:

- Ofereix la possibilitat de treballar amb DataFrames (el que equivaldria a una matriu), objecte propi d'aquesta llibreria.
- Incorpora eines de lectura i escriptura de dades de formats típics com fitxers .csv o .txt, així com fulles de càlcul com .xslx.
- Facilitat en la fusió i unió de dades, així com en la remodelació i el pivoteig.
- Possibilitat d'addició i transformació de dades amb la funció group by, fen possible l'aplicació de diferents operacions en conjunts de dades específics que presenten alguna característica en comú, la qual els permet agrupar-los fàcilment.

La preparació i la transformació de les dades del projecte es realitzarà mitjançant l'ús d'aquesta llibreria.

2.3.3. Spyder / Anaconda

Anaconda és una distribució de Python formada per un gran nombre de paquets de llibreries i entorns de treball de ciència de dades. Aquesta permet la instal·lació i posada en marxa d'aplicacions i editors com Spyder, i proporciona una interfície gràfica des de la qual es poden manipular els diferents entorns. Del paquet que ofereix Anaconda en la seva instal·lació, es farà gran ús de l'IDE Spyder i de la llibreria Scikit-learn, entre d'altres.

El codi aplicat durant el treball s'editarà en el programari *Spyder,* un IDE (*Integrated Development Environment)* desenvolupat en Python, el qual consisteix principalment en una aplicació informàtica que proporciona un marc de treball en el desenvolupament del software. Els principals blocs de treball són:

- **Editor:** L'editor multi-llenguatge integra diferents eines per tal de garantir una experiència òptima en l'edició. Algunes de les eines són la detecció de possibles errors en el codi i la seva conseqüent alerta, la explicació de funcions i classes i el seu fàcil accés a les seves corresponents definicions o l'anàlisi del codi en temps real.

- **Consola**: La consola integrada és concretament *IPython,* que permet executar comandes i interactuar amb el sistema de forma gràfica. Facilita la posada en marxa de fitxers i la interrupció d'execucions sense afectar altres elements de *Spyder*.

- **Explorador de variables**: L'explorador de variables mostra tots els elements creats durant la sessió, és a dir, mostra les variables, funcions o mòduls creats. Permet també la eliminació directa de variables o la seva edició.

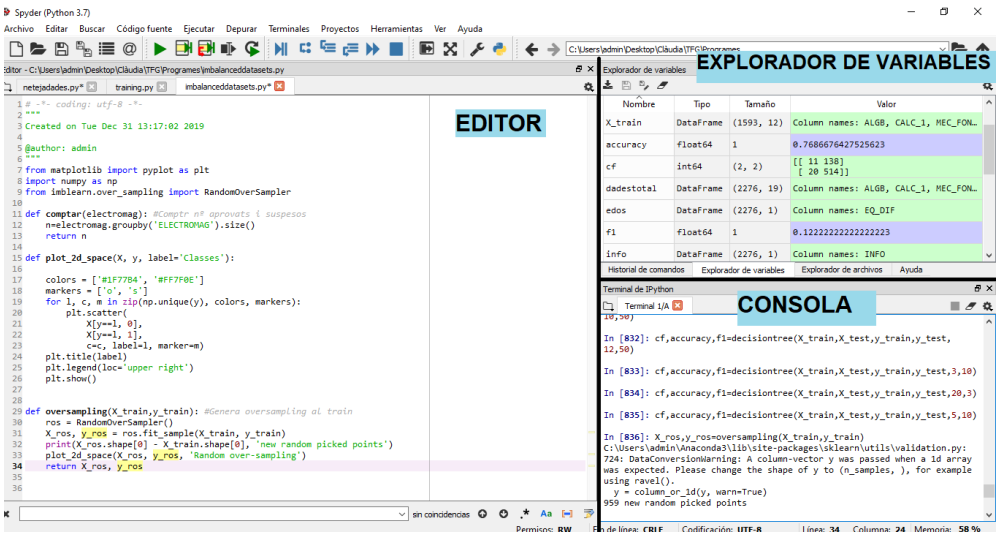


Figura 1: *Blocs de treball de Spyder i la seva situació en la pantalla*

2.3.4. Scikit-learn

Scikit-learn és una llibreria que ve ja instal·lada amb la distribució Anaconda. Aquesta llibreria proporciona diversos algorismes que faciliten l'anàlisi de dades, proporcionant així un estalvi de temps en la programació de codi per a la construcció del model i permeten facilitar el seu posterior anàlisi.

3. LA MINERIA DE DADES

3.1. Definició del concepte

En termes generals, entenem com a mineria el procés d'extraure material de valor de la Terra, com seria la mineria del carbó o del diamant. En el cas de la mineria de dades (o Data Mining en anglès), es refereix a l'extracció d'informació útil on la font és un conjunt molt gran de dades. Pel que fa al Data Mining, el resultat de l'extracció no són dades en si, sinó un conjunt de patrons i coneixement que s'obté al final d'aquest procés, amb la finalitat d'obtenir conclusions que col·laborin en la millora del projecte.

L'objectiu principal de la mineria de dades és recollir informació que ajudi a preveure patrons ocults, tendències futures i comportaments que poden ajudar a prendre decisions en el projecte. Tècnicament, el Data Mining és el procés computacional d'analitzar dades des de diferents perspectives i dimensions, establint diferents categories per arribar a la informació significant.

En el camp de la mineria de dades, es coneix com a **Mineria de dades educativa** aquella que analitza les dades existents en un camp educatiu, incloent la interacció dels usuaris amb l'escenari i els resultats d'aquests usuaris. Té el mateix objectiu que les analítiques d'aprenentatge: establir una millora en la pràctica educativa, una de les finalitats principals d'aquest projecte.

3.1.1. Procés

Per a dur a terme el procés de la mineria de dades, cal realitzar-ho en diferents fases o etapes:

1. Integració del conjunt de dades: en primer lloc, tota la informació provinent de les diferents fonts es reunida i integrada en un sol arxiu o document.

2. Selecció del conjunt de dades: es realitza una tria d'aquelles dades que es considera que seran útils per a la finalitat del projecte.

3. Filtre del conjunt de dades: poden existir errors, valors buits o dades inconsistents (que entenem com a "soroll"). En aquesta fase cal retirar-los del nostre conjunt de dades a analitzar.

4. Transformació del conjunt de dades d'entrada: es realitza amb l'objectiu de preparar el conjunt per aplicar la tècnica de mineria de dades que millor s'adapti, és a dir, es

fa un preprocessament de les dades.

5. Selecció i aplicació de la tècnica de mineria de dades: construcció i aplicació del model predictiu.

6. Avaluació de patrons i extracció de coneixement: mitjançant el model de coneixement obtingut, observació dels patrons de comportament a través de les variables del problema i les relacions d'associació entre aquestes variables.

7. Interpretació de les dades: realitzar una validació del model comprovant que les conclusions extretes són vàlides i posteriorment fer ús del nou coneixement obtingut per a la presa de decisions.

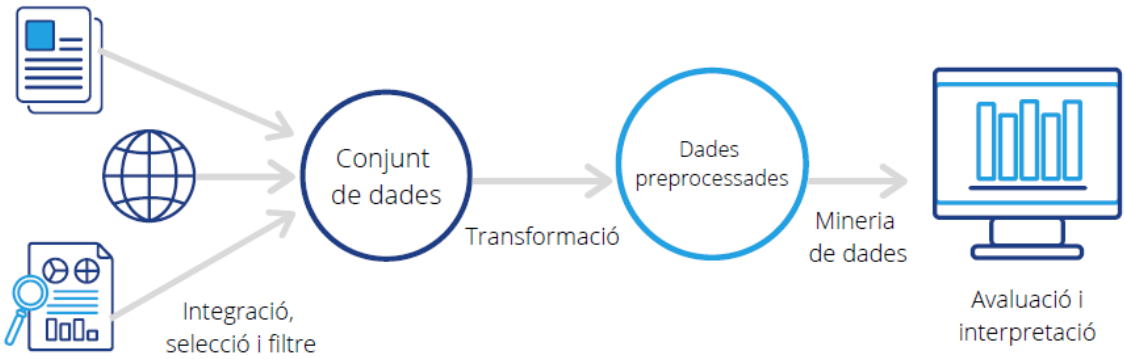


Figura 2: *Procés de la mineria de dades*

3.1.2. Aplicacions

La tècnica de la mineria de dades per a processar grans quantitats d'informació té moltes aplicacions quotidianes, per exemple:

- Millora de l'assistència sanitària: la mineria de dades té un gran potencial per a la millora dels sistemes sanitaris. S'utilitza les dades i anàlisi per a identificar aquelles pràctiques que milloren l'assistència i en redueixen el cost. Per exemple, es pot utilitzar la mineria de dades per preveure el volum de pacients en una categoria i així reduir el temps d'espera en els centres d'atenció sanitària.

- Anàlisi del cistell de compra: aquesta tècnica permet al comerciant entendre el comportament d'un client, ja que es preveu que si el client adquireix un cert grup de productes, es pot establir un altre grup de productes en què el client estigui interessat.

Aquesta informació permet als comerços adaptar el seu negoci d'acord amb les necessitats o interessos dels clients que els freqüenten, utilitzant comparacions entre diferents comerços en diferents grups demogràfics.

- Enginyeria de sistemes de fabricació: la mineria de dades pot ser utilitzada per a establir diferents relacions entre l'estructura d'un producte, el seu cost, el seu temps de fabricació, etc. i així poder optimitzar-ne el procés d'obtenció.

- Sistemes CRM (Customer Relationship Management): podem utilitzar tècniques de Data Mining per tal de millorar la lleialtat del client respecte al comerç, per a la implementació d'estratègies orientades al client ...

- Detecció de frau: els mètodes tradicionals per a la detecció de frau solen ser complexos i requereixen molt de temps. La mineria de dades proporciona patrons significants i permet treballar amb dades d'interès, ja que aquestes han estat prèviament filtrades. Això permet agilitzar el procés i preveure més fàcilment si una acció és fraudulenta o no.

- Detecció d'intrusions: mitjançant la millora de sistemes per evitar intrusions en sistemes (com seria l'autentificació de l'usuari, la protecció d'informació, etc.), es pot reduir considerablement el nombre d'intrusions mitjançant la detecció d'anomalies.

- Bioinformàtica: el procés de mineria de dades pot ser molt útil per a la bioinformàtica, ja que és molt ric en dades. Minar dades biològiques permet extreure informació significant de bases molt gran de dades, per tal d'obtenir millores en camps com la medicina o la neurociència. Algunes aplicacions serien el mapeig del genoma, el diagnòstic de malalties, la optimització del tractament de malaltes...

- Educació: com s'ha esmentat anteriorment, la mineria de dades educativa vol preveure el comportament acadèmic dels estudiants, els efectes d'un suport educacional i fer avenços en el coneixement científic sobre l'aprenentatge. Amb l'ajuda dels resultats que proporciona aquesta tècnica, les institucions poden realitzar millores en els continguts a ensenyar però també a la manera en què s'ensenyen.

3.2. CRISP-DM

CRISP DM (Cross Industry Standard Process for Data Mining) és un model de mineria de dades que ens proporciona una descripció normalitzada del cicle de vida d'un projecte estàndard d'anàlisi de dades. El model CRISP-DM cobreix les fases d'un projecte, amb les

seves respectives tasques i les relacions entre aquestes.

La metodologia CRISP-DM contempla un procés d'anàlisi de dades com un projecte professional, establint així un context molt més ric que influeix en la elaboració dels models. Aquest context contempla l'existència d'un client extern al equip de desenvolupament, així com el fet de que un projecte no s'acaba quan s'arriba al model idoni, ja que aquest requereix un manteniment posterior.

La metodologia CRISP DM es pot explicar mitjançant sis fases. La seqüència d'aquestes fases no és rígida, sinó que es pot descriure de forma cíclica.

3.2.1. Comprensió del negoci

Inclou la determinació dels objectius del negoci, l'avaluació de la situació actual, l'establiment dels objectius de la mineria de dades i el desenvolupament d'un pla de projecte.

3.2.2. Comprensió del conjunt de dades

Una vegada els objectius del negoci i el pla de projecte estan establerts, la comprensió de les dades contempla els requeriments d'aquestes. Aquesta fase pot incloure la reunió inicial de dades, la descripció i exploració d'aquestes i finalment una verificació de la seva qualitat. Al final d'aquesta fase, es pot donar l'exploració de les dades mitjançant eines estadístiques (que inclou la visualització categòrica de les variables).

3.2.3. Preparació de les dades

Una vegada identificades les fonts que ens poden proporcionar dades, aquestes necessiten ser seleccionades, filtrades, estructurades de la forma desitjada i posades en el format adequat. En aquesta fase s'ha de dur a terme el filtre i la transformació de les dades necessàriament. D'altra banda, també es pot donar una exploració de les dades en més profunditat i poden ser emprats models addicionals, permetent així identificar patrons basats en la comprensió del negoci.

3.2.4. Modelatge

Per a l'anàlisi inicial, les eines de software de mineria de dades (com seria la representació gràfica, l'establiment de relacions entre variables o un anàlisi per grups) poden ser útils. Una vegada s'ha adquirit un major coneixement de les dades (sovint a través de la identificació de patrons), poden ser aplicats models més detallats adequats al tipus de dades.

3.2.5. Avaluació i validació

En aquesta fase del projecte, s'ha construït un o varis models que pretenen obtenir qualitat suficient des de la perspectiva d'anàlisi de dades. Abans de procedir al desplegament final del model, cal avaluar-lo rigorosament i revisar els passos realitzats per crear-lo, però també comparar el model obtingut amb els objectius del negoci i comprovar que aquests han estat considerats suficientment. Al final d'aquesta fase, s'hauria d'obtenir una decisió sobre l'aplicació dels resultats del procés d'anàlisi de dades.

3.2.6. Implementació

La mineria de dades pot ser emprada tant per verificar hipòtesis prèvies com per obtenir nous coneixements (la identificació de noves i útils relacions).

Generalment, la creació del model no suposa el final del projecte, ja que el coneixement obtingut probablement s'haurà d'organitzar i presentar al client per a que aquest pugui fer-ne ús.

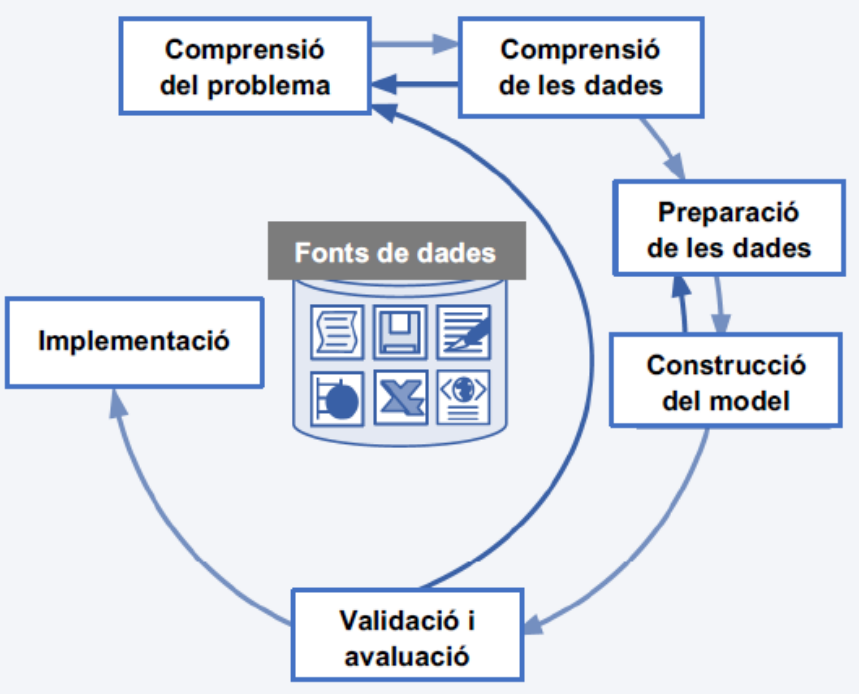


Figura 3: *Representació cíclica de les fases de CRISP*

Segons el portal KDNuggets, especialitzat en ciència de dades, CRISP és la metodologia més utilitzada en projectes de mineria de dades, imposant-se amb un 43% d'ús respecte a les altres tecnologies.

Aquesta metodologia reflecteix les fases que seran realitzades per dur a terme aquest projecte. El problema presentat consisteix en la voluntat per a predir el comportament acadèmic en el seu tercer quadrimestre dels estudiants del Grau en Enginyeria en Tecnologies Industrials de la ETSEIB, mitjançant les notes que aquests han obtingut en la fase inicial del grau, és a dir, durant el primer i segon quadrimestre.

La principal problemàtica que presenta el projecte és la facilitat que representa la predicció d'un aprovat respecte a la dificultat de la predicció d'un suspès. El fet que en les dades originals hi hagi una descompensació gran entre els aprovats i els suspesos (ja que n'hi ha molts més d ela primera classe), fa que el model de predicció no predigui correctament els suspesos, per tant aquest no s'ajusta correctament a la realitat.

Seguint la metodologia CRISP, es pretén al llarg del projecte poder trobar una solució al problema inicialment plantejat.

4. COMPRENSIÓ I PREPARACIÓ DE LES DADES

Una de les primeres fases a realitzar per tal de poder obtenir un bon resultat per a l'estudi, és la comprensió i la preparació de les dades inicials.

4.1. Dades inicials

Inicialment, es disposa d'un conjunt d'informació de tots dels estudiants de l'Escola Tècnica d'Enginyeria Industrial de Barcelona que han accedit a l'Escola entre els anys 2010 i 2017. Es disposa de tres fitxers de format Microsoft Excel que contenen la informació detallada a continuació.

El primer fitxer (*dadesnomespreins*) inclou informació del o la estudiant en el moment de la seva inscripció a la universitat, abans de cursar-hi cap assignatura. La informació es donada a partir de les següents variables:

Variable	Definició
CODI_EXPEDIENT	Codi de sis xifres adjudicat a un o una estudiant per tractar-ne el seu expedient.
SEXE	Indica si l'estudiant és home (H) o dona (D).
CP_FAMILIAR	Codi postal del lloc de residència de l'alumne.
ANY_ACCES	Any en què l'estudiant accedeix a l'Escola.
TIPUS_ACCES	Tipus d'accés a l'Escola, és un valor indicat com a 1.
NOTA_ACCES	Nota que l'estudiant va obtenir en les Proves d'Accés a la Universitat.
CP_CENTRE_SEC	Codi postal del lloc on es troba el centre en què l'estudiant va cursar la seva educació prèvia.

Taula 1: *Columnes de dadesnomespreins*

El segon fitxer (*qfaseini*) inclou informació sobre l'estudiant en la seva fase inicial, és a dir, de les assignatures que pertanyen al primer i al segon quadrimestre del grau. Les variables utilitzades són:

Variable	Definició
CODI_EXPEDIENT	Codi de sis xifres adjudicat a un o una estudiant per tractar-ne el seu expedient.
CODI_PROGRAMA	Indica quin grau cursa l'estudiant. En aquest cas, el Grau en Enginyeria en Tecnologies Industrials correspon al 752.
CODI_UPC_UD	Codi adjudicat a l'assignatura de la qual en disposem la seva qualificació.
CREDITS	Nombre de crèdits que té l'assignatura.
CURS	Any en què s'està cursant l'assignatura.
QUAD	Quadrimestre en què s'ha cursat l'assignatura, s'indica com a Q1 el quadrimestre cursat a la tardor i com a Q2 el quadrimestre cursat a la primavera.
SUPERA	Variable binària que indica si l'estudiant aprova (S) o no aprova (N) l'assignatura.
NOTA_PROF	Nota final que el professor adjudica a l'estudiant.
NOTA_NUM_AVAL	Nota obtinguda en l'assignatura posteriorment a l'avaluació. curricular.
NOTA_NUM_DEF	Nota final definitiva obtinguda quan tanca el període d'avaluació curricular.
GRUP_CLASSE	Grup de classe en què estava matriculat l'estudiant al cursar l'assignatura.

Taula 2: *Columnes de qfaseini i qfasenoini*

Finalment, es disposa d'un tercer fitxer (*qfasenoini*) que indica les mateixes variables que les indicades anteriorment, però en aquest cas de les assignatures que no pertanyen a la fase inicial, sinó a la resta del grau.

A partir d'aquestes dades inicials donades, caldrà aplicar un cert conjunt de transformacions per tal de facilitar-ne la seva exploració i així obtenir-ne el màxim d'informació possible.

4.2. Preparació de les dades

El tractament de les dades serà mitjançant la llibreria Pandas, per tant els tres arxius que contenen taules explicats anteriorment, seran convertits a format DataFrame. Inicialment caldrà convertir les taules de format Excel (com venen donades) a format Csv, per tal de poder operar amb elles.

4.2.1. Transformació de les dades

En primer lloc, s'ha fet un filtre a les dades donades per tal de seleccionar únicament aquelles que són de interès per a la predicció. El programa escrit en què s'indiquen les funcions es troba en l'Annex I.

En el fitxer *dadesnomespreins* s'ha eliminat les columnes SEXE, CP_FAMILIAR, ANY_ACCES, TIPUS_ACCES i CP_CENTRE_SEC, quedant així la nota obtinguda a les PAU. Posteriorment, s'ha eliminat del fitxer aquells alumnes que no hagin superat la Fase Inicial (és a dir, el seu CODI_EXPEDIENT no apareix a *qfasenoini*).

En els fitxers *qfasein*i i *qfasenoini*:

- S'ha seleccionat només aquells alumnes que cursin el Grau en Enginyeria en Tecnologies Industrials (codi 752).
- S'ha prescindit d'aquells alumnes que han convalidat l'assignatura (GRUP_CLASSE apareix com a CONV).
- S'ha eliminat les columnes CODI_PROGRAMA, CREDITS, GRUP_CLASSE, SUPERA, NOTA_PROF, NOTA_NUM_AVAL.
- S'ha seleccionat únicament les assignatures que pertanyen al primer, segon i tercer quadrimestre (Q1,Q2 i Q3).
- S'ha eliminat de *qfaseini* aquells alumnes que no hagin superat la fase inicial, és a dir, els que no apareixen a *qfasenoini.*
- Per aquells alumnes que hagin cursat més d'una vegada una assignatura de la fase inicial, és a dir, que hagin necessitat més d'una convocatòria per aprovar-la, s'ha seleccionat la seva darrera nota i la resta han estat eliminades.
- Per aquells alumnes que hagin cursat més d'una vegada una assignatura del Q3, s'ha seleccionat la seva primera nota d'entre totes les convocatòries realitzades, ja sigui aprovada o no, ja que és la variable que es pretén predir.

4.2.2. Nom de les assignatures segons el seu Codi UPC:

Un altre dels canvis aplicats ha estat el canvi del codi UPC per un codi alfabètic que indica el nom de l'assignatura, per tal d'agilitzar-ne l'estudi. A la taula següent es troba la equivalència entre el nom de l'assignatura i el seu corresponent codi segons la Universitat. La tercera columna indica a quin quadrimestre del grau pertany, essent 1 i 2 la Fase Inicial i 3 el primer quadrimestre de segon any. La quarta columna indica el codi alfabètic imposat.

Assignatura	Codi UPC	Quadrimestre	Codi Alfabètic
Càlcul I	240012	1	CALC_1
Àlgebra	240011	1	ALGB
Fonaments d'Informàtica	240015	1	FON_INFO
Mecànica Fonamental	240013	1	MEC_FON
Química I	240014	1	QUIM_1
Càlcul II	240022	2	CALC_2
Expressió Gràfica	240025	2	EXPRE
Termodinàmica Fonamental	240023	2	TERMO_FON
Química II	240024	2	QUIM_2
Geometria	240021	2	GEOM
Mètodes Numèrics	240032	3	MET_NUM
Informàtica	240132	3	INFO
Equacions Diferencials	240131	3	EQ_DIF
Mecànica	240133	3	MEC
Electromagnetisme	240031	3	ELECTROMAG
Materials	240033	3	MAT

Taula 3: *Noms de les assignatures i corresponent codi UPC*

4.2.3. Funció pivot

Posteriorment a la neteja de les dades, s'obtenen els fitxers *qfaseini* i *qfasenoini* modificats, on tenim a cada fila una convocatòria d'una assignatura d'un determinat expedient. Es desitja crear una nova taula, on totes les dades d'un mateix alumne (és a dir, un codi d'expedient) es trobin en una única fila.

Mitjançant Pandas, cal fer ús de la funció *pivot*, que fa que cada fila del DataFrame inicial (*qfaseini o qfasenoini*) passi a ser una columna del DataFrame final *(tfaseini o tfasenoini*). En el DataFrame *tfaseini* o *tfasenoini,* cada fila correspon a un alumne i cada columna pertany a una assignatura de la fase inicial o del tercer quadrimestre, respectivament. El valor que es troba dins la cel·la és la nota numèrica que va obtenir l'estudiant en l'assignatura corresponent.

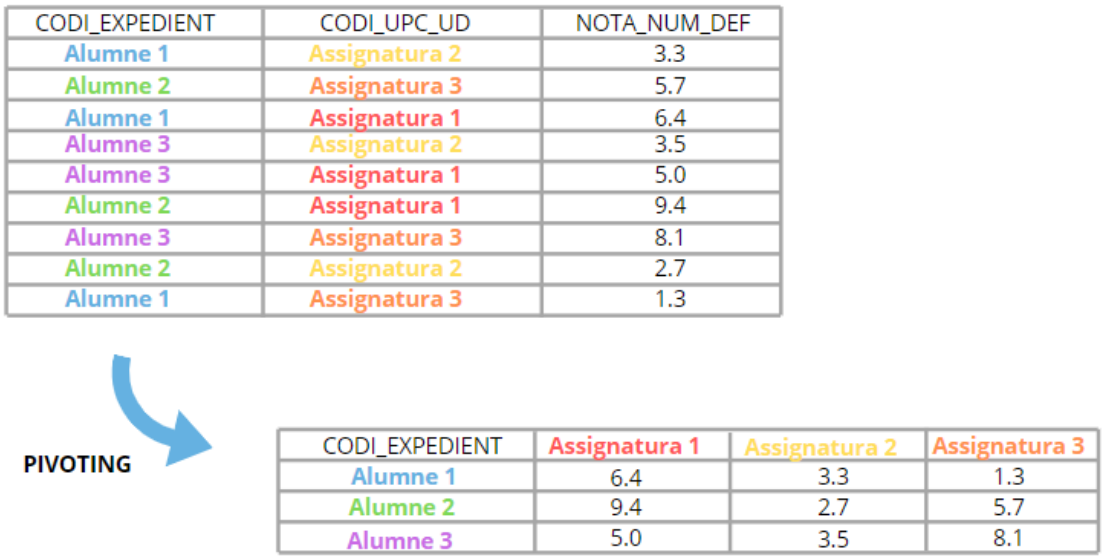


Figura 4: *Procés del pivoting*

Finalment, es crea un DataFrame (*dadestotal*) que és la unió dels dos anteriors, per tant, cada fila conté totes les notes de Q1,Q2 i Q3 d'un o una alumne.

4.2.4. Creació de noves columnes

Per tal de poder prosseguir amb l'estudi, cal que creem noves columnes per tal de facilitar un posterior modelatge. En la següent taula es detalla quines han estat les columnes afegides i la seva funció.

Variable	Definició
NOTA_ACCES	Nota amb què l'estudiant va accedir a l'Escola. Prové del fitxer *dadesnomespreins.*
MITJANA_FI	Nota mitjana de l'alumne de totes les assignatures de la Fase Inicial.
MITJANA_FNI	Nota mitjana de l'alumne de les assignatures del tercer quadrimestre.

Taula 4: *Noves columnes afegides a dadestotal*

5. MODELATGE

Un cop obtingudes les dades degudament estructurades i preparades, es procedeix al seu anàlisi. Com bé indica la metodologia Crisp-DM, en primer lloc fer la construcció del model i posteriorment fer-ne la seva validació. Aquestes dues fases van íntimament lligades, ja que per tal de construir un model cal validar un model inicial, posteriorment modificat a partir dels resultats obtinguts, ja que és un procés cíclic.

5.1. Tipus de mecanismes:

Per tal de realitzar l'anàlisi de les dades, cal disposar de dades etiquetades per classes. És a dir, en aquest treball es pretén predir el valor d'un atribut (qualificació d'una assignatura del tercer quadrimestre) el qual depèn del valor d'altres atributs (assignatures de la fase inicial).

Es distingeix, doncs, tres grans tipus d'anàlisi:

- **Classificació**: El model prediu mitjançant la identificació de la classe a la que pertany un objecte. S'utilitza, per exemple, en la detecció de correus no desitjats (*spam*) o en el reconeixement d'imatges. Aplicat en aquest projecte, seria la predicció de l'aprovat o el suspès d'un determinat estudiant en una determinada assignatura.

- **Regressió:** Predicció de l'atribut continu d'un objecte. S'empra, per exemple, en el càlcul dels preus de l'estoc. Aplicat en aquest projecte, com que en aquest cas es prediu valors numèrics (no categòrics, com en la classificació), seria la predicció de la nota que obtindrà l'estudiant en la determinada assignatura, essent un valor entre 0 i 10.

- **Clusterització** (*clustering*): Consisteix en la partició de les dades en subconjunts similars. Un exemple seria la segmentació en l'estudi de perfils de clients en comerços.

Aquest projecte se centrarà en l'ús de mecanismes de classificació. Aquests permeten una visualització més clara tant de les dades com dels resultats obtinguts. Concretament, en aquest projecte s'emprarà els arbres de decisió per a dur a terme el modelatge. Els algorismes emprats s'han obtingut de la llibreria Scikit-Learn (o sk-learn), de manera que ja estan preparats per a ser utilitzats directament.

5.2. Arbres de decisió

Els arbres de decisió són diagrames que presenten una estructura, com bé indica el seu nom, en forma d'arbre. El diagrama parteix d'una arrel i seguidament s'apliquen una sèrie de condicions. L'acompliment o no d'aquestes és el que elabora diferents camins que condueixen finalment a una resposta.

Un arbre de decisió està composat per dos elements principals: els nodes i les branques.

Un node (o fulla) representa un atribut que posteriorment es subdividirà en altres nodes, en funció de les condicions aplicades. Existeixen tres tipus de nodes:

- Node arrel: És el primer element des d'on s'inicia el diagrama, que s'acaba desglossant en la resta de nodes.
- Nodes terminals: Són els que es troben al final de l'arbre. Representa l'atribut resposta a la qüestió plantejada.
- Nodes interns: Són tots aquells que es troben entre el node arrel i els nodes terminals. Tots ells es subdivideixen en altres nodes i provenen d'un node superior, en funció de les condicions.

D'altra banda, les branques representen les unions entre els nodes. Una branca surt d'un únic node i entre en un altre node. La seva procedència i direcció depèn de la resposta que s'hagi obtingut de l'atribut del seu node procedent.

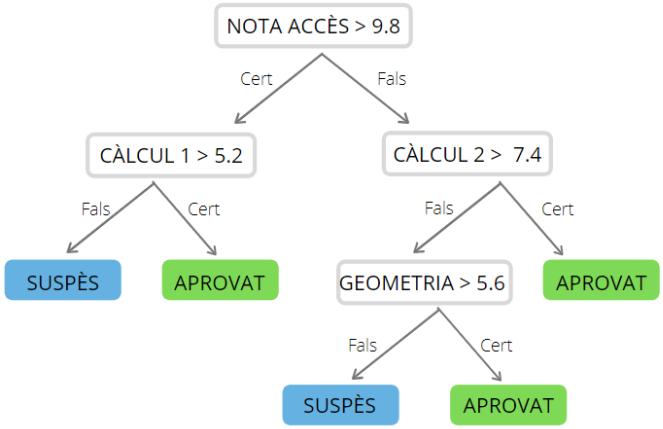


Figura 5: *Funcionament dels arbres de decisió (fictici)*

En la següent figura s'il·lustra un exemple d'arbre de decisió. Aquest prediu l'aprovat o suspès (en aquest cas fictici) de l'assignatura d'Equacions Diferencials. Com es pot veure, el node arrel en aquest cas és la Nota d'Accés a la Universitat de l'estudiant. Seguidament, s'indica que en funció de les notes que l'estudiant hagi obtingut a les assignatures de Càlcul 1, Càlcul 2 i Geometria (nodes interns), aquest aprovarà o no (nodes terminals) l'assignatura de Equacions diferencials. Les fletxes entre nodes representen les branques.

L'algorisme, per tal de construir un arbre, es basa en l'anàlisi de dades per subconjunts. Es realitzen diferents subconjunts dins del conjunt total de dades aplicant condicions fins que totes les dades que formen el subconjunt prenen el mateix valor en la categoria que es vol predir. S'empra, per tant, el recurs de la recursivitat: el conjunt es va dividint en subconjunts fins que s'arriba a un subconjunt pur.

5.2.1. Funcionament de l'algorisme

Per tal d'obtenir els resultats abans descrits, s'empra la funció DecisionTreeClassifier de la llibreria Scikit-Learn. Aquesta funció admet les dades d'entrada en forma numèrica (Nota decimal obtinguda en l'assignatura) , però la predicció que realitza es binària (Aprovat/Suspès), és per això que l'assignatura a predir s'haurà de passar prèviament a Aprovat / Suspès abans d'introduir-la a l'algorisme.

Cal remarcar que l'algorisme DecisionTreeClassifier no admet Missing Values o Nan, és per això que és molt important que en el filtratge i neteja prèviament realitzat al conjunt de dades, aquest tipus de valors s'eliminin.

L'algorisme mesura les desigualtats mitjançant l'eina *gini*, que mesura la qualitat del tall realitzat. És a dir, *gini* és el criteri que utilitza l'arbre de decisió per a decidir quins nodes s'utilitzen per a realitzar talls en el conjunt de dades donades. Gini és doncs un paràmetre matemàtic que calcula la probabilitat de no trobar el valor correcte d'una classe després d'un node, per tant, pren 0 quan ens trobem en un subconjunt pur (i assolim, per tant, un node terminal), ja que ja s'ha trobat el valor correcte en tots els subconjunts de dades.

D'entre tots els paràmetres que poden ser definits quan s'empra l'algorisme de predicció DecisionTreeClassifier, en destaquem tres:

- **Random State**: Valor enter que representa la llavor emprada pel generador de nombres aleatoris. En aquest projecte s'emprarà sempre random_state=0.

- **Min_samples_leaf**: Defineix el nombre mínim de mostres que ha de contenir un node per a ser considerat com a tal. El punt de tall a una deguda profunditat, només serà considerat si deixa un subconjunt amb min_samples_leaf mostres dins els subconjunts que creen tant la branca dreta com l'esquerra. Pot ser un nombre enter (int) o un nombre decimal (float).

- **Max_depth**: Profunditat màxima a la que pot arribar l'arbre. Si no s'introdueix cap valor (o sigui, per defecte) l'algorisme expandeix els nodes fins que s'arriba a un subconjunt pur o fins que el subconjunt té menys mostres que *min_samples_leaf*.

6. VALIDACIÓ

6.1. Mètodes de validació

Per tal de saber la precisió del model construït anteriorment, cal realitzar una validació d'aquest. Posteriorment, es realitzaran una sèrie de canvis en el model predictiu en funció del resultat d'aquesta validació inicial, per a millorar-ne la precisió. Finalment, es realitzarà una validació final.

Aprendre a fer el model de predicció i validar-lo amb les mateixes dades, però, suposa un problema. Això és degut a que el model repetiria els atributs de les mostres que just acaba de validar i s'obtindria doncs, una puntuació perfecta. El problema, però, seria que quan s'introduïssin noves dades, aquest no les prediria bé. Aquest fenomen s'anomena *overfitting*. Per tal d'evitar-ho, és comú utilitzar un algorisme que ens permeti retenir part de les dades disponibles com a dades d'entrenament (*training*) i l'altra part restant s'empri per a realitzar la validació (*testing*).

Dues de les estratègies més populars de validació que permeten evitar l'*overfitting* són:

- **Holdout** Consisteix en la subdivisió de les dades en dos conjunts: el *training* i el *testing*. El conjunt de dades de *training* s'emprarà per a construir el model, mentre que el conjunt de *testing* s'utilitzarà per validar aquest.

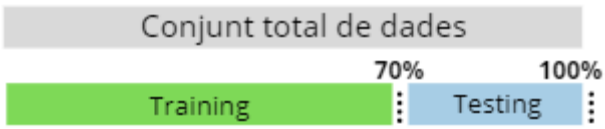


Figura 6: *Funcionament del holdout*

- **K-Cross-Validation** Es divideix les dades en k subconjunts, d'entre els quals un es fa servir únicament com a *testing*. La resta de conjunts (k-1), s'utilitzen per al *training.* Aquests conjunts, però, es subdivideixen en k-1 subconjunts, dels quals un d'ells s'emprarà també per al *testing*. La figura següent il·lustra el procés per a k=4.

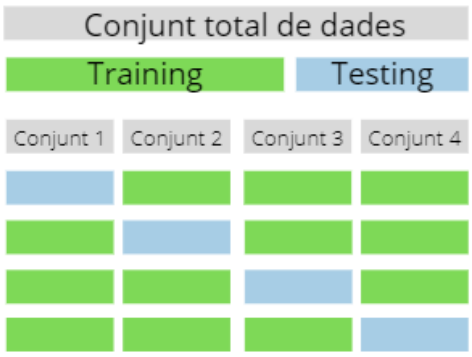


Figura7: *Funcionament de K-Cross-Validation*

Per aquest projecte, s'ha decidit emprar la metodologia de *Holdout,* retenint un 70% de les dades com a *training* i un 30% com a *testing*. Això dóna un total de 1365 expedients per al *training* i 911 expedients per al *testing*.

6.1.1. *Accuracy*

L'*accuracy* (en català, precisió) és una de les eines emprades en el procés de validació. Indica la fracció de prediccions correctes obtingudes respecte al nombre de prediccions realitzades. Matemàticament, es defineix com:

$$\texttt{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

On y circumflexa és la predicció realitzada de la variable y i $n_{samples}$ és el nombre total de mostres.

6.1.2. Matriu de confusió

Dins del procés de validació, també s'ha emprat una eina que no és un mètode de validació en si, però que ens proveeix informació rellevant: la matriu de confusió. La matriu ens serà útil per a preveure quin tipus d'error és més comú, és a dir, si el model prediu millor els aprovats o els suspesos. Aplicat al projecte:

SC (Suspès cert): Predicció correcta d'un suspès.

FS (Fals suspès): Es prediu com a suspès però és aprovat.

FA (Fals aprovat): Es prediu com aprovat però és suspès.

AC (Aprovat cert): Predicció correcta d'un aprovat.

Figura 8: *Matriu de confusió*

En un model amb 100% d'encerts, FS i FA serien igual a 0. Com més alts siguin FA i FS en relació a SC i AC, pitjor es la precisió del model, ja que hi ha més quantitat de prediccions incorrectes.

6.1.3. Eina F1

Una altra de les eines emprades en el procés de validació és F1. Per tal de definir-la correctament, cal definir dos conceptes: ***Precision*** i ***Recall***.

- ***Precision:*** Analitza la relació entre els suspesos reals i els suspesos predits-

$$\text{Precision} = \frac{\text{Suspès cert}}{\text{Suspès cert + Fals suspès}} = \frac{\text{Suspès cert}}{\text{Total suspesos predits}}$$

- Recall: La eina *recall* calcula quants dels suspesos predits són suspesos en realitat.

$$\text{Recall} = \frac{\text{Suspès cert}}{\text{Suspès cert + Fals aprovat}} = \frac{\text{Suspès cert}}{\text{Total suspesos reals}}$$

La eina F1 és funció de *precision* i *recall.* Aquesta eina serveix per trobar un balanç entre les dues eines anteriors. La diferència entre la puntuació obtinguda amb F1 i la obtenció obtinguda amb accuracy, és que l'eina accuracy es veu altament influenciada per un gran nombre de Falsos Suspesos, cosa que no ens interessa. La eina F1 pot servir d'ajuda per a trobar un equilibri entre el nombre d'aprovats i de suspesos. Es defineix com:

$$F1 = 2 \times \frac{\text{precision x recall}}{\text{precision + recall}}$$

6.2. Predicció mitjançant arbres de decisió

L'anàlisi predictiu de dades es realitza aplicant algorismes basats en arbres de decisió. Com ja s'ha descrit anteriorment, els arbres de decisió presenten dos paràmetres principals que han de ser definits: min_sample_leaf i max_depth. La construcció d'arbres de decisió es fa aplicant diversos valors d'aquests dos paràmetres, i es calcula la precisió que cada combinació d'ells dóna mitjançant en mètode de valuació emprat. El rang de valors de max_depth es troba entre 3 i 20, mentre que min_samples_leaf es troba entre 1 i 50. Les precisions indicades són en tant per 1, per tant el valor 1.00 correspon a un predictor ideal.

6.2.1. Electromagnetisme

S'aplica el model de predicció anteriorment descrit a l'assignatura de Electromagnetisme. Mitjançant diferents combinacions dels paràmetres esmentats (max_depth i min_sample_leaf), s'obté l'*accuracy* per a cada una d'elles. Els resultats es mostren a la taula 5.

S'observa que en un valor fixe de min_samples_leaf, la precisió del model disminueix a mesura que va creixent el paràmetre max_depth. Aquest fet era previsible, ja que com més s'augmenta la profunditat de l'arbre, més s'ajusten les dades al mecanisme de predicció i pitjor precisió s'obté. També s'observa que quan es fixa un valor de max_depth, la precisió del model tendeix a augmentar a mesura que va creixent el valor de min_sample_leaf, ja que s' imposar un valor alt de mostres per node.

A la taula, s'ha ressaltat la combinació dels paràmetres que aconsegueix una major precisió, la qual pren de valor 0,6939. Aquesta combinació s'obté per a un valor de min_sample_leaf de 50, mentre que sembla que el valor de max_depth no influeix molt en el resultat, ja que s'obté la mateixa precisió per a 7,10,12,14,16,18 i 20.

Accuracy		Max_depth								
		3	5	7	10	12	14	16	18	20
Min_sample_Leaf	1	0.6896	0.6647	0.6545	0.6515	0.6471	0.6252	0.6339	0.6413	0.6547
	3	0.6896	0.6691	0.6706	0.6662	0.6574	0.6749	0.6574	0.6662	0.6662
	5	0.6896	0.6589	0.6603	0.6691	0.6574	0.6559	0.6691	0.6720	0.6633
	10	0.6896	0.6530	0.6647	0.6515	0.6471	0.6530	0.6471	0.6530	0.6530
	15	0.6896	0.6603	0.6749	0.6618	0.6589	0.6647	0.6636	0.6589	0.6618
	20	0.6896	0.6896	0.6706	0.6896	0.6896	0.6896	0.6896	0.6896	0.6896
	25	0.6896	0.6705	0.6779	0.6808	0.6808	0.6808	0.6808	0.6808	0.6808
	30	0.6896	0.6647	0.6881	0.6911	0.6911	0.6911	0.6911	0.6911	0.6911
	35	0.6896	0.6647	0.6896	0.6955	0.6955	0.6955	0.6955	0.6955	0.6955
	40	0.6896	0.6647	0.6896	0.6896	0.6896	0.6896	0.6896	0.6896	0.6896
	50	0.6896	0.6764	0.6939	0.6939	0.6939	0.6939	0.6939	0.6939	0.6939

Taula 5: *Valor d'accuracy per a Electromagnetisme*

Posteriorment, s'ha analitzat amb l'eina F1 (calculant primerament precision i recall). Els resultats obtinguts es troben a la taula 6.

S'observa que succeeix, en general, el mateix que amb l'accuracy. En fixar un valor de

max_depth, s'observa que F1 augmenta a mesura que anem augmentant min_sample_leaf.

Quan fixem un valor de min_sample_leaf i anem augmentant max_depth, s'observa que en alguns casos F1 augmenta i posteriorment disminueix, cosa que no succeïa amb l'accuracy. De totes maneres, s'observa que el major valor de F1 (0.5451, marcat en blau) s'obté per a combinacions grans de max_depth i min_sample_leaf.

F1		Max_depth								
		3	5	7	10	12	14	16	18	20
Min_sample_Leaf	1	0.4879	0.4242	0.4562	0.4968	0.4869	0.4968	0.4948	0.4877	0.4886
	3	0.4879	0.4293	0.4738	0.4979	0.5106	0.5359	0.5188	0.5063	0.5105
	5	0.4879	0.4218	0.4834	0.4865	0.5193	0.5149	0.5188	0.5182	0.5084
	10	0.4879	0.3999	0.4899	0.4739	0.4859	0.4859	0.4859	0.4749	0.4859
	15	0.4879	0.4751	0.5022	0.5000	0.5043	0.4957	0.5011	0.5032	0.4978
	20	0.4879	0.4779	0.5213	0.5331	0.5331	0.5331	0.5331	0.5331	0.5331
	25	0.4879	0.4779	0.5089	0.5261	0.5261	0.5261	0.5261	0.5261	0.5261
	30	0.4879	0.4455	0.5103	0.5279	0.5279	0.5279	0.5279	0.5279	0.5279
	35	0.4879	0.4455	0.5330	0.5517	0.5517	0.5517	0.5517	0.5517	0.5517
	40	0.4879	0.4455	0.5450	0.5451	0.5451	0.5451	0.5451	0.5451	0.5451
	50	0.4879	0.5143	0.5345	0.5345	0.5345	0.5345	0.5345	0.5345	0.5345

Taula 6: *Valor d'F1 per a Electromagnetisme*

Es pren com a solució òptima del model, doncs, la combinació min_sample_leaf= 40 i com que max_depth pren un valor entre 10 i 20, es pren max_depth=10. La matriu de confusió obtinguda s'observa a continuació.

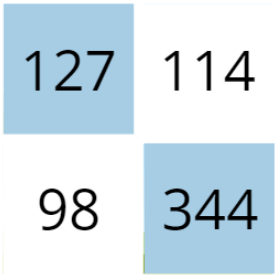


Figura 9: *Matriu de confusió per a md=10 i msl=40.*

La matriu indica que 127 suspesos han estat predits correctament, 344 aprovats han estat predits correctament, 144 suspesos han estat predits com aprovats i 98 aprovats han estat predits com a suspesos.

Una vegada realitzades les avaluacions corresponents, s'obté doncs:

Accuracy				F1			
Max		Min		Max		Min	
0.6939		0.6647		0.5451		0.4242	
msl	md	msl	md	msl	md	msl	md
50	10-20	1	5	40	10-20	1	5

Taula 7: *Resum dels resultats obtinguts per a Electromagnetisme*

On Max indica el valor màxim, Min indica el valor mínim, msl és min_sample_leaf i md és max_depth.

6.2.2. Mecànica

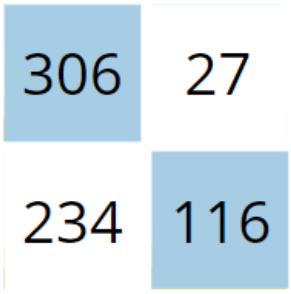
Una vegada aplicat el model de predicció a l'assignatura de Mecànica, s'obtenen els resultats indicats a les taules que es troben en els annexos. Mitjançant diferents combinació dels paràmetres esmentats (max_depth i min_sample_leaf), s'obté l'*accuracy* i l'F1 per a cada una d'elles. A continuació es troba una taula resum que remarca els valors més importants obtinguts:

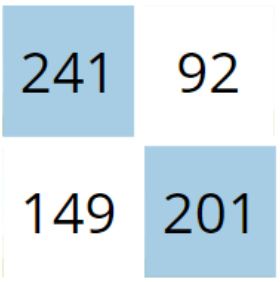
Accuracy				F1			
Max		Min		Max		Min	
0.6471		0.5622		0.7005		0.5771	
msl	md	msl	md	msl	md	msl	md
50	7-20	5	14	1-50	3	5	14

Taula 8: *Resum dels resultats obtinguts per a Mecànica*

De la mateixa manera que en l'assignatura d'Electromagnetisme, tant l'*accuracy* com l'F1 tendeixen a empitjorar quan s'augmenta el valor de max_depth i tendeixen a millorar (excepte en alguns casos) quan s'augmenta la variable min_sample_leaf.

S'observa que en aquest cas la mesura F1 obté una millor predicció en què la combinació max_depth és 3 i min_sample_leaf varia entre 1 i 50. En canvi, l'accuracy òptim es troba per a una combinació de max_depth que varia entre 7 i 20 i quan min_sample_leaf pren de valor 50. A continuació es mostren les matrius de confusió per a les combinacions md=3/msl=50 i md=10/msl=50.





Figures 10 i 11: *Matrius de confusió per a md=3/msl=50 i per a md=10/msl=50.*

S'observa que per a la combinació òptima de F1 (3/50) s'obté un major nombre de suspesos ben predits, 306, que en la combinació d'*accuracy* (10-50), que en prediu 241. Donat que el projecte es centra en la optimització de predicció de suspesos, es considera

que el fet de predir més suspesos, tot i que això comporti una menor predicció correcta d'aprovats (116 en lloc de 201) és positiu. Per tant, consideraríem la combinació md=3/msl=50 millor a l'alternativa presentada.

6.2.3. Informàtica

Els resultats obtinguts de l'aplicació del model es troben als annexos. D'entre les combinacions plantejades de min_sample_leaf i max_depth, s'obté uns valors màxims i mínims que es troben en la taula següent:

Accuracy				F1			
Max		Min		Max		Min	
0.7877		0.7218		0.3309		0.1117	
msl	md	msl	md	msl	md	msl	md
1-50	3-5	1	16	3	20	15	7

Taula 9: *Resum dels resultats obtinguts per a Informàtica*

Com es pot observar a la taula d'*accuracy*, aquesta mesura empitjora en augmentar el valor de max_depth per un mateix valor de min_sample_leaf. Pel que fa a la variació de min_sample_leaf per a un valor de max_depth, l'*accuracy* generalment tendeix a millorar.

S'observa que una de les combinacions que dona el resultat òptim és per a min_sample_leaf igual a 3 i un max_depth de 20.

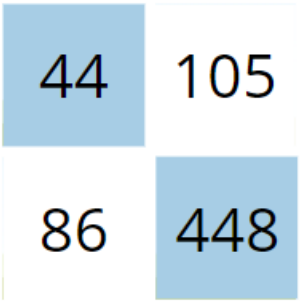


Figura 12: *Matriu de confusió per a md=20 i msl=3.*

Com bé mostra la matriu, 44 suspesos i 448 aprovats han estat predits correctament. 105 suspesos han estat predits com aprovats i 86 aprovats han estat predits com a suspesos. En comparació a assignatures com Mecànica o Electromagnetisme, s'observa que la predicció de suspesos és més aviat baixa, en relació a la predicció dels aprovats.

6.2.4. Equacions diferencials

Mitjançant l'aplicació del model i variant els paràmetres min_sample_leaf i max_depth, s'ha aconseguit elaborar les taules que es troben als annexos. A continuació es mostra una taula resum amb els valors més significatius:

Accuracy				F1			
Max		Min		Max		Min	
0.7862		0.7071		0.2966		0.0869	
msl	md	msl	md	msl	md	msl	md
1-25	3-5	10	18	5	20	30	3-5

Taula 10: *Resum dels resultats obtinguts per a Equacions Diferencials*

S'observa que en aquest cas el paràmetre F1 és molt baix, bastant inferior al que s'ha obtingut en les assignatures analitzades anteriorment. En canvi, l'accuracy obté valors molt elevats. Això és degut a que F1 pondera amb més importància la predicció dels suspesos que la predicció dels aprovats, mentre que l'accuracy actua al revés. Es pot deduir, per tant, que el model prediu correctament els aprovats mentre que amb els suspesos no és tant eficient. A continuació es mostra la matriu de confusió d'una de les combinacions òptimes, és a dir, que dóna més suspesos certs:

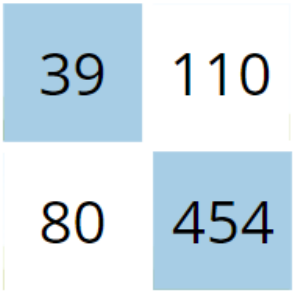


Figura 13: *Matriu de confusió per a md=20 i msl=5.*

Com es pot observar, només 39 suspesos s'han predit correctament, un número molt baix en comparació als aprovats certs, 454. Això pot ser degut a que el DataFrame que proporciona les dades per a l'ajust del model, conté molts aprovats però pocs suspesos, cosa que fa que la predicció d'aquest últim no sigui molt bona.

6.2.5. Mètodes numèrics

La taula amb les precisions obtingudes es troba en els annexos. A continuació, la taula resum:

Accuracy				F1			
Max		Min		Max		Min	
0.8712		0.7906		0.2595		0.1176	
msl	md	msl	md	msl	md	msl	md
25-35	1-20	3	14-16	5	14	1-20	3-20

Taula 11: *Resum dels resultats obtinguts per a Mètodes Numèrics*

Tal i com succeïa amb l'assignatura d'Equacions Diferencials, a Mètodes Numèrics tampoc s'obté una F1 molt alta en comparació a l'accuracy. Altre cop això s'atribueix al fet que el DataFrame qfasenoini no conté un gran nombre de suspesos en relació al nombre d'aprovats i això fa que el model no predigui bé la classe minoritària.

La matriu de confusió obtinguda per a la combinació min_sample_leaf igual a 5 i una max_depth de 14 és:

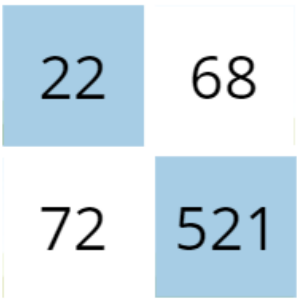


Figura 14: *Matriu de confusió per a md=14 i msl=5.*

Es confirma doncs, que la predicció de suspesos encertats (22) és molt menor a la predicció d'aprovats certs (521).

Cal comentar que en algunes combinacions de F1 s'obté resultat infinit, i això és degut a que s'obté un nombre de suspesos certs igual a zero, i això causa la divisió entre un quocient nul.

6.2.6. Materials

Finalment, per a les diferents combinacions de max_depth i min_sample_leaf, s'obté les taules amb els valors de F1 i Accuracy que es troben en els annexos. A continuació es mostra la taula resum:

Accuracy				F1			
Max		Min		Max		Min	
0.6939		0.5915		0.4499		0.2678	
msl	md	msl	md	msl	md	msl	md
1	3	1	14	5	16	5	7

Taula 12: *Resum dels resultats obtinguts per a Materials*

En aquest cas s'observa que tot i que la predicció encertada dels suspesos és inferior a la predicció correcta dels aprovats, no presenta un desequilibri tant gran com en Equacions Diferencials o Mètodes Numèrics. La matriu de confusió per a max_depth igual a 16 i un min_sample_leaf de 5 mostra un total de 99 suspesos ben predits i 334 aprovats encertats.

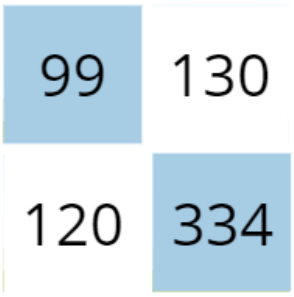


Figura 15: *Matriu de confusió per a md=16 i msl=5.*

7. TÈCNIQUES DE MOSTREIG PER A IMBALANCED DATASETS

7.1. Imbalanced datasets

Un conjunt de dades no equilibrat (*imbalanced data set*), és un conjunt de dades que en la predicció d'una variable (en aquest cas, l'aprovat o suspès en una assignatura), presenta molt més d'un valor que de l'altre. En el data set del projecte, hi ha una clara diferència en el nombre d'aprovats que de suspesos, cosa que dificulta la predicció encertada d'aquest últim.

Assignatura	Nº Aprovats	Nº Suspesos	Proporció
Electromagnetisme	1069	524	2,04:1
Mecànica	882	771	1,14:1
Informàtica	1276	317	4,03:1
Eq.Diferencials	1299	294	4,42:1
Mètodes Numèrics	1389	204	6,81:1
Materials	1110	483	2,29:1

Taula 13: *Demostració del desequilibri entre els aprovats i els suspesos en y_train*

És per això que es procedeix a aplicar tècniques de mostreig per a reduir el desbalanç entre el nombre de punts d'una classe i l'altra, per exemple, el *resampling* (o re-mostreig).

7.2. Resampling

Una de les tècniques emprades per a tractor un conjunt de dades molt desequilibrat, és el resampling, és a dir, tornar a fer el mostreig. Consisteix en eliminar mostres de la classe majoritària (under-sampling) o en afegir més mostres de la classe minoritària (over-sampling).

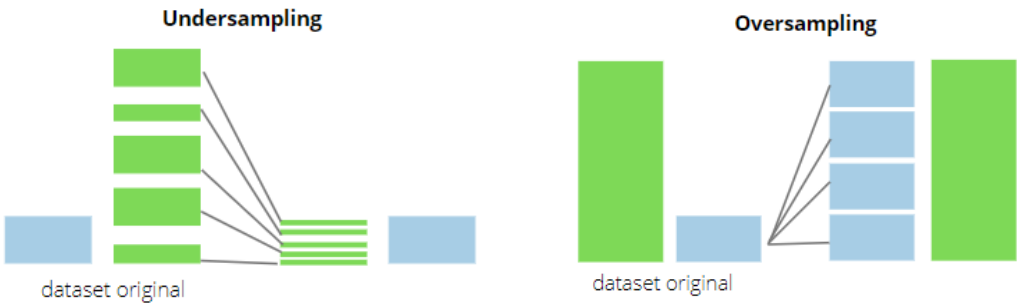


Figura 16: *Processos d'undersampling (esquerra) i oversampling (dreta).*

Tot i els avantatges que aquestes mesures presenten, també poden presentar febleses. La implementació més simple d'*oversampling* consisteix en duplicar valors aleatoris de la classe minoritària (és a dir, els suspesos), la qual cosa pot causar *overfitting*. En el cas de l'*undersampling*, la tècnica més simple implica eliminar mostres aleatòries que pertanyen a la classe majoritària, la qual cosa pot portar a una pèrdua important d'informació.

En aquest projecte s'implementarà la tècnica de l'*oversampling* aleatori. Per tal de poder executar-ho en *python*, s'ha emprat la llibreria **imblearn** (*imbalanced learn*), que ofereix un nombre de tècniques per a fer *resampling* en conjunts de dades que presenten una gran diferència entre classes. És compatible amb les funcions de *scikit-learn*.

La funció emprada per a generar els nous DataFrames (amb nous punts generats gràcies al *random oversampling*) s'anomena ***ovesampling*** i es troba als annexos.

7.3. Predicció amb arbres de decisió després de l'*oversampling*

Una vegada aplicada la tècnica de l'*oversampling*, es recalcula la precisió que presenta el model de predicció per a cada assignatura i emprant diferents combinacions de *max_depth* i *min_sample_leaf*.

7.3.1. Electromagnetisme

Mitjançant l'aplicació de l'oversampling al DataFrame destinat al training, es generen 545 punts aleatoris nous. Treballem doncs amb un DataFrame de training que conté 1069 aprovats i 1069 suspesos.

Per tal d'avaluar l'eficàcia de l'oversampling, es torna a calcular l'accuracy i F1 de l'assignatura d'electromagnetisme. Els resultats obtinguts es troben a les taules 15 i 16, respectivament.

Accuracy				F1			
Max		Min		Max		Min	
0.6735		0.6120		0.6144		0.4524	
msl	md	msl	md	msl	md	msl	md
20-25	10-20	1-50	3	50	7-20	1	18

Taula 14: *Resum dels resultats obtinguts per a Electromagnetisme post oversampling*

Una de les combinacions que proveeixen una millor predicció dels suspesos és la de min_sample_leaf igual a 50 i max_depth igual a 10. La matriu de confusió obtinguda és:

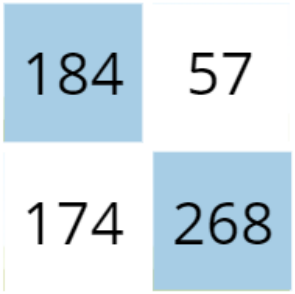


Figura 17: *Matriu de confusió per a md=10 i msl=50.*

Com es pot observar, tot i que la predicció dels suspesos certs (184) segueix essent inferior a la dels aprovats certs (268), la primera ha millorat respecte a la obtinguda amb les dades sense oversampling, tot i que això comporti una disminució en l'encert d'aprovats.

Seguidament es detallen els resultats d'Accuracy i F1 obtinguts, a les taules 15 i 16, respectivament.

Accuracy		Max_depth								
		3	5	7	10	12	14	16	18	20
Min_sample_Leaf	1	0.6120	0.6427	0.6369	0.6442	0.6354	0.6325	0.6354	0.6208	0.6296
	3	0.6120	0.6442	0.6325	0.6471	0.6325	0.6223	0.6281	0.6208	0.6164
	5	0.6120	0.6471	0.6310	0.6369	0.6281	0.6354	0.6310	0.6369	0.6354
	10	0.6120	0.6486	0.6383	0.6500	0.6515	0.6471	0.6471	0.6486	0.6471
	15	0.6120	0.6486	0.6530	0.6486	0.6486	0.6486	0.6486	0.6486	0.6486
	20	0.6120	0.6501	0.6618	0.6735	0.6735	0.6735	0.6735	0.6735	0.6735
	25	0.6120	0.6486	0.6735	0.6735	0.6735	0.6735	0.6735	0.6735	0.6735
	30	0.6120	0.6530	0.6530	0.6603	0.6603	0.6603	0.6603	0.6603	0.6603
	35	0.6120	0.6501	0.6486	0.6486	0.6486	0.6486	0.6486	0.6486	0.6486
	40	0.6120	0.6501	0.6544	0.6544	0.6544	0.6544	0.6544	0.6544	0.6544
	50	0.6120	0.6617	0.6617	0.6617	0.6617	0.6617	0.6617	0.6617	0.6617

Taula 15: *Valors d'accuracy obtinguts per a Electromagnetisme posterior al oversampling*

F1		Max_depth								
		3	5	7	10	12	14	16	18	20
Min_sample_Leaf	1	0.6039	0.5947	0.5491	0.5475	0.5257	0.4990	0.4929	0.4524	0.4718
	3	0.6039	0.5957	0.5451	0.5528	0.5219	0.4981	0.4960	0.4789	0.4781
	5	0.6039	0.5990	0.5351	0.5491	0.5171	0.5146	0.5154	0.5249	0.5221
	10	0.6039	0.5986	0.5400	0.5465	0.5352	0.5228	0.5189	0.5200	0.5189
	15	0.6039	0.6091	0.5775	0.555	0.5506	0.5506	0.5506	0.5506	0.5506
	20	0.6039	0.6101	0.5823	0.6735	0.5752	0.5752	0.5752	0.5752	0.5752
	25	0.6039	0.6078	0.5878	0.5863	0.5863	0.5863	0.5863	0.5863	0.5863
	30	0.6039	0.6069	0.5698	0.5639	0.5639	0.5639	0.5639	0.5639	0.5639
	35	0.6039	0.6049	0.5819	0.5819	0.5819	0.5819	0.5819	0.5819	0.5819
	40	0.6039	0.6049	0.5903	0.5903	0.5903	0.5903	0.5903	0.5903	0.5903
	50	0.6039	0.6138	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144

Taula 16: *Valors d'F1 obtinguts per a Electromagnetisme posterior al oversampling.*

7.3.2. Mecànica

Mitjançant la tècnica de l'oversampling, s'ha generat 51 punts nous. Aquesta mesura ens ha permès tornar a avaluar el model, els resultats obtinguts en les taules es troben als annexos. Els resultats es troben resumits a continuació:

Accuracy				F1			
Max		Min		Max		Min	
0.6589		0.5695		0.6729		0.5739	
msl	md	msl	md	msl	md	msl	md
40-50	3	5	20	35	7-20	5	20

Taula 17: *Resum dels resultats obtinguts per a Mecànica post oversampling*

S'observa que en aquest cas la mesura F1 és lleugerament superior a l'accuracy, tot i que presenten valors molt semblants. Podem deduir-ne, doncs, que obtindrem una millor predicció dels suspesos certs que dels aprovats certs. A continuació és mostra la matriu de confusió per a una de les combinacions òptimes, min_sample_leaf igual a 35 i max_depth igual a 10.

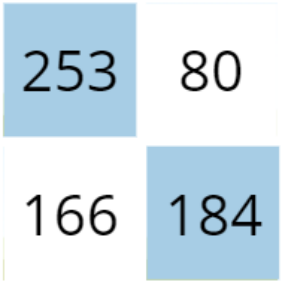


Figura 18: *Matriu de confusió per a md=10 i msl=35.*

La matriu mostra que el nombre de suspesos certs (253) és superior al nombre d'aprovats certs (184) predits. La tècnica de l'oversampling ens ha permès predir millor els suspesos a canvi de "sacrificar" una mica la predicció dels aprovats. Tot i així, es considera que l'oversampling ha millorat la predicció del model, tot i que no es considera molt precís.

7.3.3. Informàtica

Una vegada aplicat l'oversampling, s'obtenen 959 nous punts generats. A continuació es mostra una taula resum dels resultats obtinguts, que es troben als annexos.

Accuracy				F1			
Max		Min		Max		Min	
0.7247		0.6296		0.4979		0.2797	
msl	md	msl	md	msl	md	msl	md
1	18	5	7	30-50	3	1	14

Taula 18: *Resum dels resultats obtinguts per a Informàtica post oversampling*

Pel que fa a l'assignatura de Informàtica, s'observa que la mesura F1 és inferior a la mesura accuracy, cosa que significa que s'obté una millor predicció dels aprovats que dels suspesos. Tot i així, no presentarà un desequilibri tant gran com si no s'hagués aplicat l'oversampling. La matriu de confusió per a max_depth=18 i min_sample_leaf=14 és la següent:

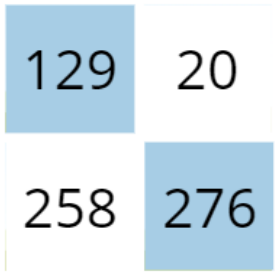


Figura 19: *Matriu de confusió per a md=18 i msl=1.*

En aquest cas s'observa que s'obté també un gran nombre de fals suspès (258), per tant el model no s'ha acabat d'ajustar bé a les dades.

7.3.4. Equacions diferencials

En el cas de l'assignatura d'Equacions Diferencials, s'observa que hi ha un gran desequilibri entre el nombre d'aprovats i el nombre de suspesos que conté el DataFrame original. Mitjançant l'aplicació de la tècnica de mostreig, s'ha generat 1005 punts nous. El resum dels resultats és el següent:

Accuracy				F1			
Max		Min		Max		Min	
0.7028		0.5461		0.4576		0.2987	
msl	md	msl	md	msl	md	msl	md
1	18	1-25	3	40	5	2	16

Taula 19: *Resum dels resultats obtinguts per a Equacions diferencials post oversampling*

S'observa que gràcies a l'oversampling, la predicció de suspesos millora notablement, ja que ara el DataFrame original conté igual nombre de suspesos i aprovats i això permet que el model s'ajusti millor. La matriu de confusió per a min_sample_leaf igual a 40 i max_depth igual a 5 és:

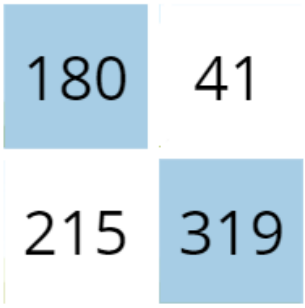


Figura 20: *Matriu de confusió per a md=5 i msl=40.*

S'observa que la predicció de suspesos certs a augmentat en comparació a l'estudi sense oversampling, però això ha causat que també augmenti el nombre de falsos suspesos (215).

7.3.5. Mètodes Numèrics

Mitjançant la tècnica de l'oversampling, s'ha obtingut una quantitat de 1185 punts nous. La taula a continuació mostra un resum dels resultats obtinguts, que es troben a l'annex.

Accuracy				F1			
Max		Min		Max		Min	
0.7906		0.5915		0.3298		0.2531	
msl	md	msl	md	msl	md	msl	md
1	20	25	5	35	12	10	16

Taula 20: *Resum dels resultats obtinguts per a Mètodes Numèrics post oversampling*

S'observa que la predicció dels suspesos segueix essent baixa en relació a la dels aprovats. Tot i així, s'ha millorat en comparació a l'estudi anterior en què no s'havia aplicat l'oversampling. La matriu de confusió per a max_depth igual a 12 i min_sample_leaf igual a 35, una de les òptimes, confirma el que s'ha comentat:

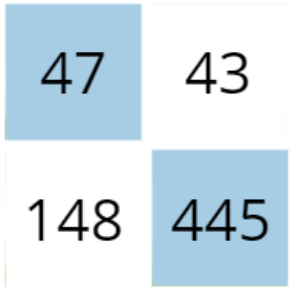


Figura 21: *Matriu de confusió per a md=12 i msl=35.*

7.3.6. Materials

La tècnica de l'oversampling genera 627 punts que tenen per objectiu igualar el nombre de suspesos i el nombre d'aprovats al DataFrame tractat. El resum dels resultats obtinguts es mostra a continuació:

Accuracy				F1			
Max		Min		Max		Min	
0.6515		0.5281		0.6003		0.3945	
msl	md	msl	md	msl	md	msl	md
50	7-20	5	16	10-50	3	1	15

Taula 21: *Resum dels resultats obtinguts per a Materials post oversampling*

Seguidament es mostra la matriu de confusió per una de les combinacions òptimes, min_sample_leaf=30 i max_depth=3:

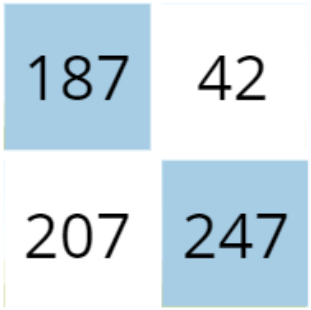


Figura 22: *Matriu de confusió per a md=3 i msl= 30.*

S'observa que en el cas de Mètodes Numèrics, la predicció correcta d'aprovats i la predicció correcta de suspesos no difereix molt, però tot i així no es considera que la predicció obtinguda pel model sigui significativa, ja que la precisió segueix essent baixa (0,6515) i F1 també (0,6003).

8. COMPARACIÓ DELS RESULTATS

	Sense oversampling		Amb oversampling	
	Valor màxim		Valor màxim	
	Accuracy	F1	Accuracy	F1
Electromagnetisme	0.6939	0.5451	0.6735	0.6144
Mecànica	0.6471	0.7005	0.6589	0.6729
Informàtica	0.7877	0.3309	0.7247	0.4979
Equacions Diferencials	0.7862	0.2966	0.7028	0.4576
Mètodes Numèrics	0.8712	0.2595	0.7906	0.3298
Materials	0.6939	0.4499	0.6515	0.6003

Taula 22: *Comparació dels resultats obtinguts tant d'accuracy com d'F1 amb i sense oversampling*

Una vegada realitzada l'avaluació del model amb i sense la tècnica d'*oversampling*, s'observa que la mesura *accuracy* empitjora, però en canvi la mesura F1 millora considerablement (en alguns casos fins i tot 0.15 més).

Aquest fet és degut a que les dades en què se'ls ha aplicat l'*oversampling* presenten més suspesos i per tant el set de dades de *training* ja no presenta un desequilibri. Això permet una millor predicció dels suspesos però una pitjor predicció dels aprovats, sacrificant el valor *accuracy* però augmentant l'F1.

Tenint en compte que l'objectiu principal d'aquest projecte és la predicció d'assignatures suspeses, aquest fet es pot considerar positiu i es considera que els resultats es confirma que les dades a les quals se'ls ha aplicat l'*oversampling* proporcionen informació més ajustada a la realitat que aquelles en què no se li ha aplicat.

9. PRESSUPOST

El cost total del present treball es pot desglossar en costos de personal i en costos d'infraestructura. A continuació es realitzen els càlculs necessaris per tal de determinar el cost final del projecte.

9.1. Costos de personal

Els costos de personal són aquells que es representen el preu del treball realitzat per un analista en el desenvolupament del projecte, ja sigui en la investigació, l'anàlisi o la presentació. En cada una d'aquestes tres fases, però, es presenta un cost diferent per hora ja que no representen la mateixa dificultat.

La investigació engloba la familiarització del problema i metodologia, fase en que s'adquireixen els coneixements necessaris per a dur a terme el projecte. Durant l'anàlisi s'executen les tasques pròpies requerides per a la resolució del problema presentat. Finalment, es presenten els resultats i se'n extreu conclusions, a més de la corresponent documentació del procés que s'ha realitzat.

9.2. Costos d'infraestructura

En aquest projecte els costos d'infraestructura engloben els recursos informàtics i el material d'oficina.

Les despeses a tenir en compte pel que fa als recursos informàtics corresponen directament a les de l'ús de l'ordenador, ja que tot el programari emprat és lliure i de codi obert, per tant no suposa cap cost.

En aquest cas, s'ha emprat un ordenador portàtil valorat en 800 euros i es considera un cost de manteniment del 10% anual del seu preu d'adquisició per un ús de 1200 hores anuals. Establint un ús de l'ordenador del 95% en total d'hores de realització del projecte, el cost de manteniment és:

$$280h \cdot 0{,}95 \cdot \frac{800€ \cdot 0{,}1}{1200h} = 17{,}75€$$

A més del cost de manteniment, en el càlcul del cost de l'ordenador cal tenir en compte el càlcul de la seva amortització. Considerant un ús de 48 setmanes a l'any durant 4 anys des

del moment de compra. L'ordenador ha estat utilitzat en el treball durant 5 mesos, és a dir, 22 setmanes. Si es descompta un dia de descans per setmana, això representa un ús final de 19 setmanes.

$$19\ setmanes \cdot \frac{\frac{800€}{4\ anys}}{48\ setmanes} = 79{,}17€$$

Per tant, els recursos informàtics representen un cost total de:

$$79{,}17 + 17{,}75 = 96{,}92€$$

En la part que comporta el material d'oficina es computen només els costos derivats de l'ús de recursos com paper o bolígrafs durant les tres fases del projecte, conjuntament amb els possibles residus generats. Es considera un cost aproximat de 20 euros.

Tal i com es mostra a la taula següent, els costos derivats de la realització del treball i per tant el pressupost del projecte són de 10.366,92 €.

PERSONAL			
Concepte	**Preu per hora**	**Hores**	**Cost**
Investigació	30€/h	50h	1500€
Anàlisi	40€/h	200h	8000€
Presentació	25€/h	30h	750€
INFRAESTRUCTURA			
Concepte			**Cost**
Recursos informàtics			96,92€
Material d'oficina			20€
IMPORT TOTAL DEL PROJECTE			10.366,92 euros

Taula 23: *Import total del projecte desglossat*

10. IMPACTE AMBIENTAL

L'impacte ambiental que representa aquest projecte és mínim, ja que es tracta d'un treball informàtic, en què les tasques han estat dutes a terme amb ordenador i no s'ha generat residus.

Pel que fa al material d'oficina, s'ha generat residus en forma de paper que no es contemplen aquest apartat ja que s'han avaluat econòmicament en l'apartat anterior, el pressupost.

Finalment, es pot tenir en compte l'ús d'energia elèctrica que consumeix l'alimentació de l'ordenador. També es pot considerar l'energia que consumeix l'encaminador (o *router*) que proporciona connexió sense fils a internet. L'enllumenat que requereix el lloc de treball també es pot tenir en compte, tot i que la majoria de tasques s'han realitzat en entorns amb llum natural o entorns en què la il·luminació era necessària tot i no estar-hi treballant.

11. PLANIFICACIÓ DEL PROJECTE

A continuació es mostra el diagrama de Gantt que representa la planificació de les diferents tasques que han estat implicades en la realització d'aquest projecte.

		Activitat	Set19	Oct19	Nov19	Des19	Gen20
INICI DEL PROJECTE	1	Definició de la metodologia					
	2	Instal·lació de les eines					
	3	Familiarització amb Pandas					
PREPARACIÓ DE DADES	4	Neteja de dades					
	5	Transformació de dades					
MODELATGE I VALIDACIÓ	6	Estudi d'algorismes de predicció					
	7	Modelatge d'algorismes					
	8	Aplicació d'algorismes i validació					
	9	Anàlisi de resultats					
CONFECCIÓ DE LA MEMÒRIA	10	Redacció de la memòria					
	11	Pressupost					
	12	Estudi impacte ambiental					
	13	Conclusions					
PRESENTACIÓ	14	Presentació del projecte					

Taula 24: *Diagrama de Gantt de la planificació del projecte*

CONCLUSIONS

Una vegada finalitzat el present treball, es considera que s'han cobert els objectius marcats inicialment. Al llarg de tot el procés d'estudi s'ha seguit la metodologia CRISP descrita inicialment, adaptada a les característiques del treball, definint les tasques a realitzar en cada fase i garantint la possibilitat de que sigui replicada a partir de la documentació proporcionada.

Pel que fa a l'anàlisi dels resultats, s'ha pogut estudiar la precisió del model de predicció emprat, l'arbre de decisió, en funció de diferents paràmetres. La validació dels models s'ha realitzat de manera sistemàtica i aplicant les mateixes condicions, per tal que els resultats puguin ser contrastats. Mitjançant l'*accuracy* i el paràmetre F1, amb el suport de les matrius de confusió, s'ha pogut comprovar que la definició dels paràmetres permet evitar el sobreajustament dels models sobre les dades. Pel que fa a la predicció dels suspesos, s'ha obtingut una precisió baixa, tot i que la posterior aplicació de la tècnica *oversampling* ha suposat una millora en aquesta.

Es considera, però, que les dades utilitzades no són prou representatives per poder ser predites, fet que s'atribueix a l'origen de les dades. Es considera que l'ús de dades purament acadèmiques corresponents l'ETSEIB no són suficients per poder predir qualificacions d'assignatures. La introducció de noves dades d'origen extern, com seria el rendiment acadèmic abans d'entrar a l'Escola o dades relacionades amb l'àmbit familiar, podrien ajudar a que els models s'ajustessin a les dades i conseqüentment millorar la precisió de la predicció.

Tot i no haver obtingut unes precisions o F1 elevats en la predicció, es considera que s'ha complert l'objectiu principal, el qual és l'estudi del rendiment de les tècniques de mineria de dades en la predicció de l'aprovat o suspès en les assignatures corresponents al tercer quadrimestre del grau.

BIBLIOGRAFIA

[1] Han, Jiawei et al. *Data Mining: Concepts and Techniques*. 3rd ed. Waltham: Elsevier 2012. 673 p. ISBN 978-0-12-381479-1.

[2] Witten, Ian H. et al. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd ed. Burlington: Morgan Kaufmann Publishers, 2011. 617 p. ISBN 978-0-12-374856-0.

[3] Olson, David L. et al. *Advanced Data Mining Techniques.* Berlin: Springer, 2008. 169 p. ISBN 978-3-54-076916-3.

[4] Pyle, Dorian. *Data Preparation for Data Mining*. San Francisco: Morgan Kaufmann Publishers, 1999. 460 p. ISBN 978 -1-55-860529-9.

[5] *Diversos autors*. *KDNuggets.* Adreça web: https://www.kdnuggets.com

[6] *Python Data Analysis Library* (documentació de la llibreria *Pandas*. Adreça web: https://pandas.pydata.org

[7] *Scikit-learn: Machine learning in Python* (documentació de la llibreria *scikit-learn*). Adreça web: https://scikit-learn.org/stable/

[8] Diversos autors. *Stackoverflow* (fòrum de desenvolupadors de codi). Adreça web: https://stackoverflow.com

[9] *Koo Ping Shung. Towards Data Science* (F1/Precision/Recall). Adreça web: https://towardsdatascience.com

[10] *Rafael Alencar. Kaggle* (Resampling strategies for imbalanced datasets). Adreça web: https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets

ANNEX

A1. Preparació de les dades

```
import pandas as pd

def netejapreins(dataframe):   #Neteja el df dadesnomespreins
      dataframe=dataframe.drop(columns=['SEXE','CP_FAMILIAR','ANY_ACCES','TIPUS_A
      CCES','CP_CENTRE_SEC'])
      return dataframe


def netejadf(dataframe):   #Neteja faseini i fasenoini
   #Seleccionar només industrials
   dataframe=dataframe[dataframe["CODI_PROGRAMA"].isin([752])]
   #Eliminar alumnes convalidats
   dataframe=dataframe[~dataframe["GRUP_CLASSE"].isin(["CONV"])]
   #Eliminar dades que no són d'interés
    dataframe=dataframe.drop(columns=["CODI_PROGRAMA","CREDITS","GRUP_CLASS
    E","SUPERA","NOTA_PROF","NOTA_NUM_AVAL"])
   #Seleccionar assignatures només de Q1 Q2 i Q3
     assignaturesprimer=          ["240011","240011","240012","240013","240014","240015",
    "240015","240021","240022","240023","240024","240025"]
    assignaturessegon=["240031","240032","240033","240131","240132","240133"]
   assignatures=assignaturesprimer + assignaturessegon
   dataframe=dataframe[dataframe["CODI_UPC_UD"].isin(assignatures)]
   #Eliminar Nan
   dataframe=dataframe.dropna(how='any')

   return dataframe
```

```
def creardataframes():

    faseini=pd.read_excel("qfaseini.xlsx")
    fasenoini=pd.read_excel("qfasenoini.xlsx")
    preins=pd.read_excel("dadespersnomespreins.xlsx")

    faseini.to_csv("faseini.csv",index=False)
    fasenoini.to_csv("fasenoini.csv",index=False)
    preins.to_csv("preins.csv",index=False)

    faseini=pd.read_csv("faseini.csv")
    fasenoini=pd.read_csv("fasenoini.csv")
    preins=pd.read_csv("preins.csv")

    #Netejar els DataFrames
    faseini=netejadf(faseini)
    fasenoini=netejadf(fasenoini)
     preins=preins.drop(columns=['CP_FAMILIAR','ANY_ACCES','TIPUS_ACCES','CP_CEN
     TRE_SEC','SEXE'])

    return faseini, fasenoini,preins


def superafaseini(faseini,fasenoini,preins):

    #Seleccionar estudiants que hagin superat la FI (formaran part d'ambdós DataFrames)
    codiexp=set(fasenoini['CODI_EXPEDIENT'])
    faseini['SUPERAT']=faseini['CODI_EXPEDIENT'].isin(codiexp).astype(int)
    faseini=faseini[faseini['SUPERAT']==1]
    faseini=faseini.drop('SUPERAT',axis=1)

    #Eliminar de preins aquells alumnes que no hagin superat la FI
    codiexp=set(fasenoini['CODI_EXPEDIENT'])
    preins['SUPERAT']=preins['CODI_EXPEDIENT'].isin(codiexp).astype(int)
    preins=preins[preins['SUPERAT']==1]
    preins=preins.drop('SUPERAT',axis=1)
    #Posar codi_exp com a índex
    preins=preins.set_index('CODI_EXPEDIENT')
```

```
  #Millor nota de cada assignatura de la FI
  faseini=faseini[faseini.NOTA_NUM_DEF>=5.0]

  return faseini,preins


def crearfasenoini(fasenoini):
  #Quedar-se únicament amb la primera nota d'entre totes les convocatòries per a cada
  assignatura de la FNI. Després crear una taula única amb expedient-fila / assignatura-
  columna / valor-nota de l'expedient a l'assignatura
  #Crear columna assignatura-convocatòria

   fasenoini=fasenoini.sort_values(by=['CURS','QUAD'])
   fasenoini['CONV']=fasenoini.groupby(['CODI_EXPEDIENT','CODI_UPC_UD']).cumcount
   ()+1
   fasenoini['ASSIGNATURA_CONVOCATORIA']=fasenoini[['CODI_UPC_UD','CONV']].ap
   ply(lambda x:'-'.join([str(x[0]),str(x[1])]),axis=1)

  electro=fasenoini[fasenoini['CODI_UPC_UD']=='240031']
  mecanica=fasenoini[fasenoini['CODI_UPC_UD']=='240133']
  info=fasenoini[fasenoini['CODI_UPC_UD']=='240132']
  edos=fasenoini[fasenoini['CODI_UPC_UD']=='240131']
  materials=fasenoini[fasenoini['CODI_UPC_UD']=='240033']
  metodes=fasenoini[fasenoini['CODI_UPC_UD']=='240032']

#Crear per assignatura una taula amb expedient com a fila i columnes les convocatories
   electro=electro.pivot(index='CODI_EXPEDIENT',values='NOTA_NUM_DEF',columns='A
   SSIGNATURA_CONVOCATORIA')
   mecanica=mecanica.pivot(index='CODI_EXPEDIENT',values='NOTA_NUM_DEF',colum
   ns='ASSIGNATURA_CONVOCATORIA')
   info=info.pivot(index='CODI_EXPEDIENT',values='NOTA_NUM_DEF',columns='ASSIG
   NATURA_CONVOCATORIA')
   edos=edos.pivot(index='CODI_EXPEDIENT',values='NOTA_NUM_DEF',columns='ASSI
   GNATURA_CONVOCATORIA')
   materials=materials.pivot(index='CODI_EXPEDIENT',values='NOTA_NUM_DEF',column
   s='ASSIGNATURA_CONVOCATORIA')
   metodes=metodes.pivot(index='CODI_EXPEDIENT',values='NOTA_NUM_DEF',columns
   ='ASSIGNATURA_CONVOCATORIA')
```

```
    #Quedar-se amb la millor nota de cada assignatura
     electro=electro['240031-1']
    mecanica=mecanica['240133-1']
    info=info['240132-1']
    edos=edos['240131-1']
    materials=materials['240033-1']
    metodes=metodes['240032-1']

     tfasenoini=pd.merge(mecanica,electro,on='CODI_EXPEDIENT',how='inner',left_index=T
     rue,)
     tfasenoini=pd.merge(tfasenoini,info,on='CODI_EXPEDIENT',how='inner',left_index=True
     ,)
     tfasenoini=pd.merge(tfasenoini,edos,on='CODI_EXPEDIENT',how='inner',left_index=Tru
     e,)
     tfasenoini=pd.merge(tfasenoini,materials,on='CODI_EXPEDIENT',how='inner',left_index
     =True,)
     tfasenoini=pd.merge(tfasenoini,metodes,on='CODI_EXPEDIENT',how='inner',left_index=
     True,)
    mitjanafni=tfasenoini.mean(axis=1)

    return tfasenoini,mitjanafni


def crearfaseini(faseini):
    #taula amb expedients com a files, assignatures FI com a columnes i valors la nota de
    l'assignatura de l'expedient
    faseini=faseini.drop(columns=["CURS","QUAD"])
     tfaseini=faseini.pivot(index='CODI_EXPEDIENT',values='NOTA_NUM_DEF',columns='C
     ODI_UPC_UD')
    tfaseini=tfaseini.dropna(how='any')
    mitjanafi=tfaseini.mean(axis=1)

    return tfaseini,mitjanafi
```

```
def canviarnomsfi(df): #canviar els codis UPC per noms de FI

    df.columns=['ALGB','CALC_1','MEC_FON','QUIM_1','FON_INFO','GEOM','CALC_2','TE
    RMO_FON','QUIM_2','EXPRE']
    mapping = {df.columns[0]:'ALGB', df.columns[1]: 'CALC_1', df.columns[2]:'MEC_FON',
    df.columns[3]: 'QUIM_1',df.columns[4]:'FON_INFO', df.columns[5]:
    'GEOM',df.columns[6]:'CALC_2', df.columns[7]: 'TERMO_FON',df.columns[8]:'QUIM_2',
    df.columns[9]: 'EXPRE'}
    df=df.rename(columns=mapping)

     return df

def canviarnomsfni(df): #canviar codis UPC per noms de FNI

    df.columns=['MEC','ELECTROMAG','INFO','EQ_DIF','MAT','MET_NUM']
     mapping = {df.columns[0]:'MEC', df.columns[1]: 'ELECTROMAG', df.columns[2]:'INFO',
     df.columns[3]: 'EQ_DIF',df.columns[4]:'MAT', df.columns[5]: 'MET_NUM'}
    df=df.rename(columns=mapping)

     return df

def taulesjuntes(tfaseini,tfasenoini):
    #Unió FI i FNI, RESULTAT: totes les notes de FI i Q3 per expedient(fila)
     dadestotal=pd.merge(tfaseini,tfasenoini,on='CODI_EXPEDIENT',how='inner',left_index=
     True,)

     return dadestotal

def mitjanes(dadestotal,mitjanafi,mitjanafni):
    #Afegir a cada expedient la seva nota mitjana de la FI i la seva nota mitjana del Q3
    dadestotal['MITJANA_FI']=mitjanafi
    dadestotal['MITJANA_FNI']=mitjanafni

    return dadestotal
```

```
def notaacces(dadestotal,preins): #Afegeix la nota d'acces de cada estudiant
    dadestotal['NOTA_ACCES']=preins['NOTA_ACCES']
    dadestotal=dadestotal.dropna(how='any')
    return dadestotal

def mitjanataules(mitjanafi,tfaseini,mitjanafni,tfasenoini,preins):
    tfaseini['MITJANA_FI']=mitjanafi
    tfaseini['NOTA_ACCES']=preins['NOTA_ACCES']
    tfasenoini['MITJANA_FNI']=mitjanafni
    tfaseini=tfaseini.dropna(how='any')
    tfasenoini=tfasenoini.dropna(how='any')
     return tfaseini,tfasenoini


def compararmides(tfaseini,tfasenoini,dadestotal):
     #Elimina possibles expedients que siguin a una taula i a l'altra no
    tfaseini=tfaseini.assign(Midaigual=tfaseini.index.isin(tfasenoini.index).astype(int))
    tfaseini=tfaseini[tfaseini["Midaigual"].isin([1])]
    tfasenoini=tfasenoini.assign(Midaigual=tfasenoini.index.isin(dadestotal.index).astype(int))
    tfasenoini=tfasenoini[tfasenoini["Midaigual"].isin([1])]
    tfaseini=tfaseini.drop(columns=['Midaigual'])
    tfasenoini=tfasenoini.drop(columns=['Midaigual'])

  return tfaseini,tfasenoini

def netejadades(): #Funció que ho executa tot sense input)
     faseini,fasenoini,preins=creardataframes()
    faseini,preins=superafaseini(faseini,fasenoini,preins)
    tfasenoini,mitjanafni=crearfasenoini(fasenoini)
    tfaseini,mitjanafi=crearfaseini(faseini)
    tfaseini=canviarnomsfi(tfaseini)
    tfasenoini=canviarnomsfni(tfasenoini)
    dadestotal=taulesjuntes(tfaseini,tfasenoini)
    dadestotal=mitjanes(dadestotal,mitjanafi,mitjanafni)
    dadestotal=notaacces(dadestotal,preins)
    tfaseini,tfasenoini=mitjanataules(mitjanafi,tfaseini,mitjanafni,tfasenoini,preins)
    tfaseini,tfasenoini=compararmides(tfaseini,tfasenoini,dadestotal)

    return dadestotal,tfaseini,tfasenoini
```

A2. Modelatge i validació

```
import pandas as pd
import sklearn
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import confusion_matrix


def binari(nota): #Passar qualificació a aprovat (A) o suspès(S)
    if nota<5:
        return 'S'
    else:
        return 'A'

def binarin(nota): #Passar qualificació a aprovat(1) o suspès(0)
    if nota=='A':
        return 1
    else:
        return 0

def taulabinari(tfasenoini): #Aplicar la funció a tots els elements del df
    tbfasenoini=tfasenoini.applymap(binari)
    return tbfasenoini
```

Implementació holdout:

```
def holdout(X,y): #Divisió dades mitjançant holdout, 70-30
    from sklearn.model_selection import train_test_split
    X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3,random_state=0)
    return X_train,X_test,y_train,y_test
```

Implementació arbres de decisió:

```
def decisiontree(X_train,X_test,y_train,y_test,max_depth,min_samples_leaf):
    from sklearn.metrics import accuracy_score
   clf=DecisionTreeClassifier(min_samples_leaf=min_samples_leaf,max_depth=max_depth)
    clf = clf.fit(X_train,y_train)
    y_pred_test = clf.predict(X_test)
    accuracy_score = accuracy_score(y_test,y_pred_test)
    cf=confusion_matrix(y_test, y_pred_test)
    precision=cf[0,0]/(cf[0,0]+cf[1,0])
    recall=cf[0,0]/(cf[0,0]+cf[0,1])
    f1=2*((precision*recall)/(precision+recall))

    return cf,accuracy_score,f1  #Obtenció matriu de confusió, accuracy i f1
```

A3. Imbalanced Datasets

```
from matplotlib import pyplot as plt
import numpy as np
from imblearn.over_sampling import RandomOverSampler

def comptar(electromag): #Comptr nº aprovats i suspesos
    n=electromag.groupby('ELECTROMAG').size()
    return n

def plot_2d_space(X, y, label='Classes'):
    colors = ['#1F77B4', '#FF7F0E']
    markers = ['o', 's']
    for l, c, m in zip(np.unique(y), colors, markers):
        plt.scatter(
            X[y==l, 0],
            X[y==l, 1],
            c=c, label=l, marker=m)
    plt.title(label)
    plt.legend(loc='upper right')
    plt.show()


    def oversampling(X_train,y_train): #Genera oversampling al train

            ros = RandomOverSampler()
            X_ros, y_ros = ros.fit_sample(X_train, y_train)
            print(X_ros.shape[0] - X_train.shape[0], 'new random picked points')
             plot_2d_space(X_ros, y_ros, 'Random over-sampling')

           return X_ros, y_ros
```

A4. PREDICCIÓ AMB ARBRES DE DECISIÓ

MECÀNICA

Accuracy		Max_depth								
		3	5	7	10	12	14	16	18	20
Min_sample_Leaf	1	0.6193	0.6310	0.6076	0.6105	0.5827	0.5856	0.5813	0.5856	0.5900
	3	0.6193	0.6266	0.6061	0.5959	0.5871	0.5769	0.5827	0.5754	0.5857
	5	0.6193	0.6193	0.6091	0.5842	0.5754	0.5622	0.5827	0.5798	0.5797
	10	0.6193	0.6442	0.6208	0.6193	0.6178	0.6193	0.6120	0.6193	0.6135
	15	0.6223	0.6413	0.6501	0.6325	0.6354	0.6368	0.6368	0.6368	0.6368
	20	0.6237	0.6457	0.6515	0.6368	0.6368	0.6368	0.6368	0.6368	0.6368
	25	0.6237	0.6457	0.6530	0.6398	0.6398	0.6398	0.6398	0.6398	0.6398
	30	0.6178	0.6398	0.6530	0.6398	0.6398	0.6398	0.6398	0.6398	0.6398
	35	0.6178	0.6398	0.6501	0.6501	0.6501	0.6501	0.6501	0.6501	0.6501
	40	0.6178	0.6398	0.6428	0.6428	0.6428	0.6428	0.6428	0.6428	0.6428
	50	0.6178	0.6427	0.6471	0.6471	0.6471	0.6471	0.6471	0.6471	0.6471

:

F1		Max_depth								
		3	5	7	10	12	14	16	18	20
Min_sample_Leaf	1	0.7005	0.6347	0.6024	0.6167	0.5946	0.5916	0.5937	0.59510	0.5965
	3	0.7004	0.6298	0.6038	0.6102	0.6061	0.5981	0.5934	0.5845	0.5939
	5	0.7004	0.6367	0.6021	0.5861	0.5821	0.5771	0.5946	0.5986	0.5906
	10	0.6941	0.6524	0.628	0.6448	0.6419	0.6409	0.6345	0.6429	0.6374
	15	0.6993	0.6485	0.6571	0.6538	0.6556	0.6565	0.6565	0.6565	0.6565
	20	0.7015	0.6493	0.6600	0.6639	0.6639	0.6639	0.6639	0.6639	0.6639
	25	0.7015	0.6493	0.6619	0.6535	0.6535	0.6535	0.6535	0.6535	0.6535
	30	0.7010	0.6496	0.6722	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667
	35	0.7010	0.6496	0.6834	0.6834	0.6834	0.6834	0.6834	0.6834	0.6834
	40	0.7010	0.6496	0.6831	0.6831	0.6831	0.6831	0.6831	0.6831	0.6831
	50	0.7010	0.6554	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667

iNFORMÀTICA

Accuracy		Max_depth								
		3	5	7	10	12	14	16	18	20
Min_sample_Leaf	1	0.7877	0.7759	0.7598	0.7365	07394	0.7321	0.7218	0.7408	0.7365
	3	0.7877	0.7745	0.7496	0.7277	0.7291	0.7174	0.7303	0.7189	0.7218
	5	0.7877	0.7716	0.7642	0.7408	0.7482	0.7379	0.7423	0.7423	0.7423
	10	0.7877	0.7877	0.7672	0.7438	0.7482	0.7379	0.7306	0.7306	0.7306
	15	0.7877	0.7877	0.7672	0.7628	0.7569	0.7496	0.7394	0.7394	0.7394
	20	0.7877	0.7877	0.7701	0.7701	0.7657	0.7687	0.7686	0.7686	0.7686
	25	0.7877	0.7877	0.7628	0.7759	0.7759	0.7759	0.7759	0.7759	0.7759
	30	0.7877	0.7877	0.7687	0.7613	0.7569	0.7569	0.7569	0.7569	0.7569
	35	0.7877	0.7877	0.7687	0.7687	0.7687	0.7687	0.7687	0.7687	0.7687
	40	0.7877	0.7877	0.7687	0.7687	0.7687	0.7687	0.7687	0.7687	0.7687
	50	0.7877	0.7877	0.7598	0.7598	0.7598	0.7598	0.7598	0.7598	0.7598

F1		Max_depth								
		3	5	7	10	12	14	16	18	20
Min_sample_Leaf	1	0.1317	0.1547	0.1800	0.2913	0.3101	0.3297	0.3357	0.3746	0.3431
	3	0.1317	0.1444	0.1739	0.2677	0.2912	0.3275	0.3285	0.3094	0.3309
	5	0.1317	0.1789	0.2222	0.3059	0.3435	0.3194	0.3623	0.3623	0.3623
	10	0.1317	0.1317	0.2011	0.2915	0.3175	0.2981	0.3134	0.3134	0.3134
	15	0.1317	0.1317	0.1117	0.1563	0.2385	0.3133	0.3101	0.3101	0.3101
	20	0.1317	0.1317	0.1798	0.2189	0.2381	0.2752	0.2752	0.2752	0.2752
	25	0.1317	0.1317	0.1981	0.2073	0.2073	0.2073	0.2073	0.2073	0.2073
	30	0.1317	0.1317	0.2617	0.2882	0.2719	0.2719	0.2719	0.2719	0.2719
	35	0.1317	0.1317	0.2617	0.2617	0.2617	0.2617	0.2617	0.2617	0.2617
	40	0.1317	0.1317	0.2617	0.2617	0.2617	0.2617	0.2617	0.2617	0.2617
	50	0.1317	0.1317	0.2931	0.2931	0.2931	0.2931	0.2931	0.2931	0.2931

EQUACIONS DIFERENCIALS

Accuracy		Max_depth								
		3	5	7	10	12	14	16	18	20
Min_sample_Leaf	1	0.7862	0.7833	0.7775	0.7321	0.7174	0.7335	0.7042	0.6984	0.6998
	3	0.7862	0.7833	0.7745	0.7204	0.7204	0.7189	0.7189	0.7116	0.7086
	5	0.7862	0.7848	0.7716	0.7379	0.7306	0.7277	0.7233	0.7174	0.7291
	10	0.7862	0.78477	0.7613	0.7218	0.7189	0.7116	0.7116	0.7071	0.7186
	15	0.7862	0.7862	0.7701	0.7643	0.7277	0.7496	0.7496	0.7496	0.7496
	20	0.7862	0.7862	0.7847	0.7847	0.7847	0.7628	0.7628	0.7628	0.7628
	25	0.7862	0.7862	0.7847	0.7877	0.7789	0.7789	0.7789	0.7789	0.7789
	30	0.7848	0.7848	0.7862	0.7862	0.7731	0.7731	0.7731	0.7731	0.7731
	35	0.7833	0.7833	0.7848	0.7848	0.7672	0.7672	0.7672	0.7672	0.7672
	40	0.7833	0.7833	0.7745	0.7745	0.7745	0.7745	0.7745	0.7745	0.7745
	50	0.7833	0.7833	0.7833	0.7783	0.7783	0.7783	0.7783	0.7783	0.7783

F1		Max_depth								
		3	5	7	10	12	14	16	18	20
Min_sample_Leaf	1	0.1098	0.1084	0.1828	0.2407	0.2249	0.2946	0.3267	0.2945	0.2907
	3	0.1098	0.1190	0.1979	0.2738	0.2682	0.3239	0.3094	0.3230	0.3162
	5	0.1098	0.1091	0.1702	0.2510	0.2580	0.2846	0.2814	0.2825	0.2966
	10	0.1098	0.1091	0.1466	0.2213	0.2441	0.2731	0.2784	0.2593	0.2657
	15	0.1098	0.1098	0.1514	0.1744	0.2249	0.2192	0.2192	0.2192	0.2192
	20	0.1098	0.1098	0.1503	0.1503	0.1503	0.2059	0.2059	0.2059	0.2059
	25	0.1098	0.1098	0.1503	0.2076	0.2705	0.2705	0.2705	0.2705	0.2705
	30	0.0869	0.0869	0.1609	0.1609	0.2289	0.2289	0.2289	0.2289	0.2289
	35	0.1294	0.1294	0.1967	0.1967	0.2535	0.2535	0.2535	0.2535	0.2535
	40	0.1294	0.1294	0.1895	0.1895	0.1895	0.1895	0.1895	0.1895	0.1895
	50	0.1294	0.1294	0.1294	0.1294	0.1294	0.1294	0.1294	0.1294	0.1294

MÈTODES NUMÈRICS

Accuracy		Max_depth								
		3	5	7	10	12	14	16	18	20
Min_sample_Leaf	1	0.8682	0.8521	0.8419	0.8199	0.8009	0.7936	0.7921	0.7936	0.8009
	3	0.8682	0.8521	0.8143	0.8126	0.8097	0.7906	0.7906	0.7862	0.7936
	5	0.8682	0.8594	0.8448	0.8199	0.8155	0.7994	0.7921	0.7950	0.7936
	10	0.8682	0.8506	0.8448	0.8287	0.8214	0.8184	0.8214	0.8184	0.8214
	15	0.8682	0.8682	0.8609	0.8507	0.8448	0.8448	0.8448	0.8448	0.8448
	20	0.8682	0.8682	0.8682	0.8682	0.8682	0.8682	0.8682	0.8682	0.8682
	25	0.8712	0.8712	0.8712	0.8712	0.8712	0.8712	0.8712	0.8712	0.8712
	30	0.8712	0.8712	0.8712	0.8712	0.8712	0.8712	0.8712	0.8712	0.8712
	35	0.8712	0.8712	0.8712	0.8712	0.8712	0.8712	0.8712	0.8712	0.8712
	40	0.8682	0.86822	0.8682	0.8682	0.8682	0.8682	0.8682	0.8682	0.8682
	50	0.8682	0.8682	0.8682	0.8682	0.8682	0.8682	0.8682	0.8682	0.8682

F1		Max_depth								
		3	5	7	10	12	14	16	18	20
Min_sample_Leaf	1	0.1176	0.1368	0.1290	0.1022	0.1500	0.1754	0.1932	0.2034	0.2093
	3	0.1176	0.1217	0.1549	0.1795	0.2262	0.2270	0.2186	0.2234	0.2209
	5	0.1176	0.1724	0.2319	0.2454	0.2588	0.2595	0.2447	0.2391	0.2378
	10	0.1176	0.1500	0.1719	0.2041	0.2078	0.2051	0.2078	0.2051	0.2078
	15	0.1176	0.1176	0.2400	0.1774	0.2089	0.2089	0.2089	0.2089	0.2089
	20	0.1176	0.1176	0.1176	0.1176	0.1176	0.1176	0.1176	0.1176	0.1176
	25	0.1698	0.1698	0.1698	0.1698	0.1698	0.1698	0.1698	0.1698	0.1698
	30	0.1698	0.1698	0.1698	0.1698	0.1698	0.1698	0.1698	0.1698	0.1698
	35	0.1698	0.1698	0.1698	0.1698	0.1698	0.1698	0.1698	0.1698	0.1698
	40	∞	∞	∞	∞	∞	∞	∞	∞	∞
	50	∞	∞	∞	∞	∞	∞	∞	∞	∞

MATERIALS

Accuracy		Max_depth								
		3	5	7	10	12	14	16	18	20
Min_sample_Leaf	1	0.6939	0.6633	0.6237	0.6179	0.6413	0.5915	0.6193	0.6208	0.6193
	3	0.6911	0.6618	0.6179	0.6222	0.6149	0.6237	0.6223	0.6120	0.6179
	5	0.6911	0.6603	0.6076	0.6354	0.6310	0.6296	0.6384	0.6384	0.6310
	10	0.6911	0.6647	0.6164	0.6281	0.6222	0.6193	0.6193	0.6164	0.6222
	15	0.6911	0.6720	0.6398	0.6325	0.6369	0.6369	0.6369	0.6384	0.6384
	20	0.6911	0.6794	0.6662	0.6691	0.6691	0.6691	0.6691	0.6691	0.6691
	25	0.6911	0.6823	0.6764	0.6794	0.6794	0.6794	0.6794	0.6794	0.6794
	30	0.6911	0.6852	0.6647	0.6676	0.6676	0.6676	0.6676	0.6676	0.6676
	35	0.6911	0.6852	0.6589	0.6706	0.6706	0.6706	0.6706	0.6706	0.6706
	40	0.6911	0.6867	0.6676	0.6794	0.6794	0.6794	0.6794	0.6794	0.6794
	50	0.6911	0.6633	0.6457	0.6457	0.6457	0.6457	0.6457	0.6457	0.6457

F1		Max_depth								
		3	5	7	10	12	14	16	18	20
Min_sample_Leaf	1	0.2818	0.3850	0.2881	0.3619	0.4289	0.3758	0.4144	0.4100	0.4298
	3	0.2699	0.3873	0.2965	0.4028	0.4063	0.4497	0.4416	0.4072	0.4314
	5	0.2699	0.3763	0.2678	0.4057	0.4167	0.4289	0.4499	0.4424	0.4273
	10	0.2699	0.3827	0.3417	0.3952	0.3915	0.3897	0.3925	0.3907	0.3915
	15	0.2699	0.4043	0.3659	0.4203	0.4233	0.4233	0.4233	0.4242	0.4242
	20	0.2699	0.4065	0.3838	0.4264	0.4264	0.4264	0.4264	0.4264	0.4264
	25	0.2699	0.3989	0.3775	0.3999	0.3999	0.3999	0.3999	0.3999	0.3999
	30	0.2699	0.3944	0.3082	0.3343	0.3343	0.3343	0.3343	0.3343	0.3343
	35	0.2699	0.3944	0.3285	0.3902	0.3902	0.3902	0.3902	0.3902	0.3902
	40	0.2699	0.4022	0.3747	0.4312	0.4312	0.4312	0.4312	0.4312	0.4312
	50	0.2699	0.3275	0.3086	0.3086	0.3086	0.3086	0.3086	0.3086	0.3086

A5. PREDICCIÓ AMB OVERSAMPLING

MECÀNICA

Accuracy	Max_depth								
Min_sample_Leaf	3	5	7	10	12	14	16	18	20
1	0.6603	0.6296	0.6325	0.5929	0.5988	0.5871	0.5871	0.6003	0.5959
3	0.6603	0.6296	0.6369	0.6018	0.6018	0.5959	0.5856	0.5915	0.5871
5	0.6603	0.6339	0.6369	0.5871	0.5695	0.5769	0.5769	0.5798	0.5695
10	0.6603	0.6413	0.6428	0.6032	0.5900	0.5974	0.5871	0.5944	0.5974
15	0.6633	0.6457	0.6544	0.6413	0.6428	0.6428	0.6428	0.6428	0.6428
20	0.6647	0.6486	0.6545	0.6428	0.6456	0.6456	0.6456	0.6456	0.6456
25	0.6647	0.6486	0.6559	0.6339	0.6339	0.6339	0.6339	0.6339	0.6339
30	0.6589	0.6428	0.6501	0.6501	0.6501	0.6501	0.6501	0.6501	0.6501
35	0.6319	0.6428	0.6398	0.6398	0.6398	0.6398	0.6398	0.6398	0.6398
40	0.6589	0.6428	0.6384	0.6398	0.6398	0.6398	0.6398	0.6398	0.6398
50	0.6589	0.6428	0.6501	0.6501	0.6501	0.6501	0.6501	0.6501	0.6501

F1	Max_depth								
Min_sample_Leaf	3	5	7	10	12	14	16	18	20
1	0.6196	0.6349	0.6367	0.6017	0.6006	0.5925	0.5901	0.6038	0.5965
3	0.6196	0.6359	0.6353	0.6058	0.6069	06091	0.5974	0.5974	0.6017
5	0.6196	0.6429	0.6353	0.5937	0.5702	0.5877	0.5818	0.5859	0.5739
10	0.6196	0.6485	0.6453	0.6134	0.6045	0.6088	0.5994	0.6082	0.6110
15	0.6278	0.6533	0.6638	0.6525	0.6611	0.6611	0.6611	0.6611	0.6611
20	0.6312	0.6552	0.6629	0.6514	0.6562	0.6562	0.6562	0.6562	0.6562
25	0.6312	0.6552	0.6609	0.6291	0.6246	0.6290	0.6246	0.6246	0.6246
30	0.6319	0.6554	0.651	0.651	0.651	0.651	0.651	0.651	0.651
35	0.6319	0.6554	0.6729	0.6729	0.6729	0.6729	0.6729	0.6729	0.6729
40	0.6319	0.6554	0.6667	0.6658	0.6658	0.6658	0.6658	0.6658	0.6658
50	0.6319	0.6554	0.6551	0.6551	0.6551	0.6551	0.6551	0.6551	0.6551

INFORMÀTICA

Accuracy		Max_depth								
		3	5	7	10	12	14	16	18	20
Min_sample_Leaf	1	0.6354	0.6384	0.6413	0.6823	0.6881	0.6984	0.7028	0.7247	0.7130
	3	0.6354	0.6369	0.6369	0.6589	0.6647	0.6691	0.6764	0.6779	0.6808
	5	0.6354	0.6369	0.6296	0.6559	0.6720	0.6691	0.6691	0.6749	0.6706
	10	0.6398	0.6457	0.6547	0.6808	0.6823	0.6969	0.6852	0.6998	0.6896
	15	0.6398	0.6383	0.6442	0.6633	0.6676	0.6676	0.6676	0.6676	0.6720
	20	0.6398	0.6398	0.6471	0.6706	0.6706	0.6706	0.6706	0.6706	0.6706
	25	0.6398	0.6471	0.6398	0.6706	0.6749	0.6706	0.6706	0.6706	0.6706
	30	0.6369	0.6530	0.6384	0.6589	0.6589	0.6589	0.6589	0.6589	0.6589
	35	0.6369	0.6545	0.6471	0.6647	0.6647	0.6647	0.6647	0.6647	0.6647
	40	0.6369	0.6545	0.6501	0.6794	0.6794	0.6794	0.6794	0.6794	0.6794
	50	0.6369	0.6457	0.6428	0.6325	0.6325	0.6325	0.6325	0.6325	0.6325

F1		Max_depth								
		3	5	7	10	12	14	16	18	20
Min_sample_Leaf	1	0.4949	0.4688	0.4662	0.3197	0.3016	0.2797	0.2927	0.3088	0.3099
	3	0.4949	0.4678	0.4723	0.3399	0.3284	0.3274	0.3283	0.3373	0.3229
	5	0.4949	0.4678	0.4628	0.3343	0.3333	0.3193	0.3314	0.3313	0.3284
	10	0.4959	0.4807	0.4732	0.4011	0.4119	0.4234	0.4173	0.4225	0.4207
	15	0.4959	0.4711	0.4439	0.3817	0.3848	0.3848	0.3848	0.3848	0.3879
	20	0.4959	0.4698	0.4459	0.3836	0.3836	0.3836	0.3836	0.3836	0.3836
	25	0.4959	0.4726	0.4509	0.4095	0.4064	0.4032	0.4032	0.4032	0.4032
	30	0.4979	0.4527	0.4523	0.4247	0.4247	0.4247	0.4247	0.4247	0.4247
	35	0.4979	0.4537	0.4609	0.4586	0.4586	0.4586	0.4586	0.4586	0.4586
	40	0.4979	0.4537	0.4605	0.4427	0.4427	0.4427	0.4427	0.4427	0.4427
	50	0.4979	0.4424	0.4352	0.4359	0.4359	0.4359	0.4359	0.4359	0.4359

EQUACIONS DIFERENCIALS

Accuracy		Max_depth								
		3	5	7	10	12	14	16	18	20
Min_sample_Leaf	1	0.5461	0.6398	0.6003	0.6428	0.6706	0.6984	0.6925	0.7028	0.6999
	3	0.5461	0.6398	0.6047	0.6325	0.6413	0.6633	0.6633	0.6647	0.6545
	5	0.5461	0.6384	0.5988	0.6266	0.6398	0.6676	0.6559	0.6691	0.6559
	10	0.5461	0.6398	0.6018	0.6486	0.6515	0.6720	0.6662	0.6735	0.6676
	15	0.5461	0.6413	0.5974	0.6749	0.6691	0.6530	0.6603	0.6574	0.6603
	20	0.5461	0.6413	0.5974	0.6310	0.6428	0.6471	0.6471	0.6471	0.6471
	25	0.5461	0.6398	0.6047	0.6310	0.6310	0.6310	0.6310	0.6603	0.6603
	30	0.5505	0.6208	0.6281	0.6457	0.6559	0.6559	0.6559	0.6559	0.6559
	35	0.5505	0.6223	0.6281	0.6633	0.6749	0.6749	0.6749	0.6749	0.6749
	40	0.5505	0.6252	0.6135	0.6398	0.6515	0.6471	0.6515	0.6471	0.6515
	50	0.5637	0.6193	0.6339	0.6589	0.6589	0.6589	0.6589	0.6589	0.6589

F1		Max_depth								
		3	5	7	10	12	14	16	18	20
Min_sample_Leaf	1	0.4404	0.4252	0.4348	0.39	0.3553	0.3268	0.3092	0.3473	0.3322
	3	0.4404	0.4279	0.4398	0.3802	0.3099	0.3072	0.2987	0.2954	0.3018
	5	0.4404	0.4242	0.4408	0.3704	0.3315	0.3058	0.3027	0.3314	0.3027
	10	0.4404	0.4306	0.4008	0.3651	0.3239	0.3373	0.3214	0.3460	0.3264
	15	0.4404	0.4394	0.4035	0.3799	0.3859	0.3844	0.3829	0.3874	0.3862
	20	0.4404	0.4394	0.4136	0.3913	0.3900	0.3899	0.3899	0.3899	0.3899
	25	0.4408	0.4332	0.4079	0.3731	0.3731	0.3731	0.3731	0.3763	0.3763
	30	0.4408	0.4478	0.4038	0.3795	0.3767	0.3767	0.3767	0.3767	0.3767
	35	0.4485	0.4439	0.3981	0.3915	0.3901	0.3901	0.3901	0.3901	0.3901
	40	0.4485	0.4576	0.3999	0.3788	0.3769	0.3868	0.3769	0.3868	0.3769
	50	0.4440	0.4468	0.4318	0.4189	0.4189	0.4189	0.4189	0.4189	0.4189

MÈTODES NUMÈRICS

Accuracy		Max_depth								
		3	5	7	10	12	14	16	18	20
Min_sample_Leaf	1	0.6662	0.6032	0.6164	0.7101	0.7408	0.7526	0.7613	0.775	0.7906
	3	0.6662	0.6032	0.6135	0.6969	0.7291	0.7452	0.7511	0.7701	0.7818
	5	0.6662	0.6002	0.6193	0.7072	0.7130	0.7262	0.7379	0.7569	0.7569
	10	0.6662	0.6018	0.6749	0.6764	0.6954	0.7247	0.7321	0.7277	0.7277
	15	0.6676	0.6032	0.6691	0.6896	0.7116	0.7408	0.7365	0.7438	0.7452
	20	0.6559	0.5974	0.6896	0.6939	0.7233	0.7321	0.7321	0.7321	0.7321
	25	0.6647	0.5915	0.6939	0.7291	0.7262	0.7482	0.7482	0.7482	0.7482
	30	0.6647	0.6018	0.7291	0.6647	0.6984	0.7042	0.7042	0.7042	0.7042
	35	0.6647	0.6149	0.6647	0.6647	0.7204	0.7101	0.7101	0.7101	0.7101
	40	0.6647	0.6149	0.6749	0.6749	0.7116	0.7116	0.7116	0.7116	0.7116
	50	0.6939	0.6354	0.7321	0.7321	0.7321	0.7321	0.7321	0.7321	0.7321

F1		Max_depth								
		3	5	7	10	12	14	16	18	20
Min_sample_Leaf	1	0.3254	0.3139	0.3212	0.3077	0.3218	0.2988	0.2756	0.2549	0.2667
	3	0.3254	0.3174	0.3265	0.3077	0.3223	0.3040	0.3032	0.2765	0.2802
	5	0.3254	0.3089	0.3264	0.3198	0.3194	0.3099	0.342	0.2966	0.2966
	10	0.3254	0.3096	0.3232	0.3242	0.2925	0.2769	0.2531	0.2560	0.2377
	15	0.3264	0.3069	0.3193	0.3205	0.2989	0.3004	0.2742	0.2857	0.2809
	20	0.3149	0.3038	0.3205	0.3193	0.3127	0.2767	0.2767	0.2767	0.2767
	25	0.3284	0.3077	0.3192	0.3273	0.2943	0.2893	0.2893	0.2893	0.2893
	30	0.3284	0.2953	0.3263	0.3082	0.3179	0.3221	0.3221	0.3221	0.3221
	35	0.3284	0.2987	0.3082	0.3082	0.3298	0.3265	0.3265	0.3265	0.3265
	40	0.3284	0.2987	0.2839	0.2839	0.2989	0.2989	0.2989	0.2989	0.2989
	50	0.3323	0.2986	0.2989	0.2989	0.2989	0.2989	0.2989	0.2989	0.2989

MATERIALS

Accuracy		Max_depth								
		3	5	7	10	12	14	16	18	20
Min_sample_Leaf	1	0.6339	0.6149	0.6120	0.6237	0.6193	0.6120	0.6135	0.6193	0.6237
	3	0.6339	0.6149	0.6135	0.6223	0.6135	0.6281	0.6252	0.6237	0.6223
	5	0.6339	0.6237	0.6193	0.6193	0.6252	0.6369	0.5281	0.6295	0.6296
	10	0.6354	0.6208	0.6193	0.6193	0.6179	0.6339	0.6252	0.6296	0.6339
	15	0.6354	0.6179	0.6223	0.6061	0.6237	0.6223	0.6252	0.6237	0.6223
	20	0.6354	0.6223	0.6398	0.6281	0.6076	0.6120	0.6120	0.6076	0.6076
	25	0.6354	0.6164	0.6384	0.6281	0.6281	0.6281	0.6281	0.6281	0.6281
	30	0.6354	0.6237	0.6398	0.6398	0.6398	0.6398	0.6398	0.6398	0.6398
	35	0.6354	0.6252	0.6486	0.6310	0.6486	0.6486	0.6486	0.6486	0.6486
	40	0.6354	0.6164	0.6310	0.6515	0.6310	0.6310	0.6310	0.6310	0.6310
	50	0.6354	0.6252	0.6515	0.6515	0.6515	0.6515	0.6515	0.6515	0.6515

F1		Max_depth								
		3	5	7	10	12	14	16	18	20
Min_sample_Leaf	1	0.5981	0.5489	0.5375	0.4870	0.4222	0.4072	0.3945	0.4064	0.4092
	3	0.5981	0.5488	0.5385	0.4756	0.4454	0.4356	0.4311	0.4251	0.4189
	5	0.5981	0.5592	0.5439	0.4881	0.4689	0.4513	0.4502	0.4651	0.4536
	10	0.6003	0.5588	0.5113	0.4737	0.4706	0.4769	0.4776	0.4762	0.4835
	15	0.6003	0.5569	0.5057	0.4817	0.4612	0.4557	0.4622	0.4567	0.4603
	20	0.6003	0.5536	0.5461	0.5348	0.4806	0.4854	0.4854	0.4806	0.4806
	25	0.6003	0.5529	0.5597	0.5512	0.5512	0.5512	0.5512	0.5512	0.5512
	30	0.6003	0.5577	0.5654	0.5654	0.5654	0.5654	0.5654	0.5654	0.5654
	35	0.6003	0.5571	0.5714	0.5516	0.5714	0.5714	0.5714	0.5714	0.5714
	40	0.6003	0.5559	0.5516	0.5719	0.5516	0.5516	0.5516	0.5516	0.55160
	50	0.6003	0.5762	0.5719	0.5719	0.5719	0.5719	0.5719	0.5719	0.5719